165 22.3.77

Psychological Bulletin

JAMES DEESE, Editor
Johns Hopkins University

Editorial Consultants

LYNN R. ANDERSON
NORMAN K. ANDERSON
DOROTHY C. ATKINS
FRANK BARRON
JOHN BARTKO
ELLIOTT BLASS
R. D. BLOCK
RICHARD BOGARTZ
JOSEPH BRADY
E. J. CAPALDI
JOHN B. CARROLL
ROSALIND CARTWRIGHT
ALPHONSE CHAPANIS
ANTHONY J. CONGER
EMORY COWAN
FRANK J. DEIDEK
WILLIAM N. DEMBER
CLINTON DESOTO
JAMES DINSMOOR
STARKEY DUNCAN
ROBERT EBEL
HOWARD EGETH
DORIS ENTWISTLE
ALBERT ERLEBACHER
JOSEPH FLEISS
NATHANIEL L. GAGE
GOLDINE GLESER
LEON GLESER
VIVIAN GOUREVITCH
MELVIN GRAVITZ
BERT GREEN
JAMES GREENO
WILLIAM W. GRINGS
ROBERT M. GUION
MASON HAIRE
DAVID HEISE
WINIFRED HILL
JERRY HIRSCH
PAUL HOFFMAN
STEWART HULSE
MARVIN IVERSON
F. H. KANFER
DANIEL P. KIMBLE
HERBERT KLAUSMEIER
EDMUND T. KLEMMER
SAMUEL J. KOMORITA
ROY LACHMAN
WALLACE LAMBERT
PETER LANG
ROBERT LEEPER
DANIEL S. LEHRMAN
ALLAN M. LEVENTHAL
MICHAEL LEWIS
DONALD B. LINDSLEY
ROBERT LINN
RICHARD A. LITTMAN
FREDERICK M. LORD
DAVID LYKKEN
MURRAY S. MIRON
WALTER MISCHEL
JAMES MYER
ARLO MYERS
ALLEN NEWELL
MERRILL NOBLE
WALTER OBRIST
DAVID OLTON
JOHN OVERALL
DAVID PALERMO
HAROLD B. PEPINSKY
E. J. PHARES
LYMAN PORTER
MALCOLM PRESTON
MILTON ROSENBAUM
IRWIN G. SARASON
WARNER K. SCHAIE
JULIUS SEEMAN
MARVIN E. SHAW
THOMAS J. SHUELL
IRWIN SILVERMAN
NORMAN SPEAR
CHARLES SPIELBERGER
JULIAN STANLEY
RALPH M. STOGDILL
BRYAN STREPP
CHARLES TART
DAVID R. THOMAS
LARRY W. THOMPSON
HARRY C. TRIANDIS
ELLIOT S. VALENSTEIN
WILLIAM E. VINACKE
JOHN VOLKMAN
ALLAN WAGNER
MICHAEL A. WALLACH
ROGER WEBB
JERRY S. WIGGINS
DAVID G. WINTER
JOHN WOODMANSEE
EDWARD ZIEGLER
JOSEPH ZUBIN

EDU ... AND PSYCHOLOGY

VOLUME 76, 1971

HELEN ORR
Managing Editor

JEANNE MCMANUS
Technical Editor

HENRY S. SALMON
Advertising and Business Manager

ANNE REDMAN
Subscription Manager

PUBLISHED MONTHLY BY
THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
1200 SEVENTEENTH ST. N. W.
WASHINGTON, D. C. 20036

$10.25

Copyright © 1971 by the American Psychological Association, Inc.

& Psyl. Research
6.3.81
J 717

CONTENTS OF VOLUME 76

Banks, William P. Confidence-Rated Recall, d_r, and Tests of Bernbach's Finite-State Theory in Recall 151
Bentler, P. M., Jackson, Douglas, N., and Messick, Samuel. Identification of Content and Style: A Two-Dimensional Interpretation of Acquiescence 186
Bernbach, Harley A. Invariance of d^* in Memory 149
Biglan, Anthony. See Mitchell, Terence R.
Block, Jack. On Further Conjectures Regarding Acquiescence 205
Brein, Michael, and David, Kenneth H. Intercultural Communication and the Adjustment of the Sojourner 215
Brigham, John C. Ethnic Stereotypes 15
Brown, Sam C. See Roenker, Daniel L.
Clark, Russell D., III. Group-Induced Shift toward Risk: A Critical Appraisal 251
Cole, Michael. See Frankel, Frederick.
David, Kenneth H. See Brein, Michael.
Davison, Gerald C. See Wilson, G. Terence.
Fiedler, Fred E. Validation and Extension of the Contingency Model of Leadership Effectiveness: A Review of Empirical Findings 128
Fleiss, Joseph L. Measuring Nominal Scale Agreement among Many Raters 378
Frankel, Frederick, and Cole, Michael. Measures of Category Clustering in Free Recall 39
Furth, Hans G. Linguistic Deficiency and Thinking: Research with Deaf Subjects 1964–1969 58
Gruver, Gene Gary. College Students as Therapeutic Agents 111
Hebb, D. O. Comment on Altruism: The Comparative Evidence 409
Hummel, Thomas J., and Sligo, Joseph R. Empirical Comparison of Univariate and Multivariate Analysis of Variance Procedures 49
Hunt, Raymond G. See Lichtman, Cary M.
Jackson, Douglas N. See Bentler, P. M.
Jones, Edward E., and Sigall, Harold. The Bogus Pipeline: A New Paradigm for Measuring Affect and Attitude 349
Jones, Mari Riess. From Probability Learning to Sequential Processing: A Critical Review 153
Kahneman, Daniel. See Tversky, Amos.
Krebs, Dennis. Infrahuman Altruism 411
Lichtman, Cary M., and Hunt, Raymond G. Personality and Organization Theory: A Review of Some Conceptual Literature 271
Light, Richard J. Measures of Response Agreement for Qualitative Data: Some Generalizations and Alternatives 365
Locke, Edwin A. Is "Behavior Therapy" Behavioristic? (An Analysis of Wolpe's Psychotherapeutic Methods) 318
Mathews, Andrew M. Psychophysiological Approaches to the Investigation of Desensitization and Related Procedures 73
McGhee, Paul E. Development of the Humor Response: A Review of the Literature 328
Messick, Samuel. See Bentler, P. M.
Mitchell, Terence R., and Biglan, Anthony. Instrumentality Theories: Current Uses in Psychology 432
Robbins, Donald. Partial Reinforcement: A Selective Review of the Alleyway Literature 415
Roenker, Daniel L., Thompson, Charles P., and Brown, Sam C. Comparison of Measures for the Estimation of Clustering in Free Recall 45
Schmidt, Richard A. Proprioception and the Timing of Motor Responses 383
Sigall, Harold. See Jones, Edward E.

SLIGO, JOSEPH R. See HUMMEL, THOMAS J.

SPRINGER, R. M. See TATE, J. D.

STRAHAN, ROBERT F. A Coefficient of Directional Correlation for Time-Series Analysis 211

TATE, J. D., AND SPRINGER, R. M. Effects of Memory Time on Successive Judgments 394

THOMPSON, CHARLES P. See ROENKER, DANIEL L.

TVERSKY, AMOS, AND KAHNEMAN, DANIEL. Belief in the Law of Small Numbers 105

VINOKUR, AMIRAM. Review and Theoretical Analysis of the Effects of Group Processes upon Individual and Group Decisions Involving Risk 231

WILKINS, WALLACE. Desensitization: Social and Cognitive Factors Underlying the Effectiveness of Wolpe's Procedure 311

WILSON, G. TERENCE, AND DAVISON, GERALD C. Processes of Fear Reduction in Systematic Desensitization 1

WINE, JERI. Test Anxiety and Direction of Attention 92

ZEDECK, SHELDON. Problems with the Use of "Moderator" Variables 295

Vol. 76, No. 1 July 1971

Psychological Bulletin

Diary No. 363
13.9.71

PROCESSES OF FEAR REDUCTION IN SYSTEMATIC DESENSITIZATION:

ANIMAL STUDIES[1]

G. TERENCE WILSON[2] AND GERALD C. DAVISON

State University of New York at Stony Brook

Research with infrahumans on fear reduction is critically reviewed with respect to the theoretical and experimental underpinnings of systematic desensitization, and several tentative conclusions are suggested: (*a*) "Reciprocal inhibition" represents an unwarranted neurologization of a highly problematical theoretical system (Hull); (*b*) the belief that conditioned fear/avoidance is largely unsusceptible to extinction procedures is inconsistent with the firm findings of experimental extinction; (*c*) the concept of "counterconditioning" is in sore need of precise, agreed-upon definition; (*d*) graded exposure to the conditioned stimulus (CS), while not consistently superior in effectiveness to repeated exposure to the full (training) intensity CS vis-à-vis extinguishing fear/avoidance, may be inferior to graded exposure paired with anxiety-competing responses; (*e*) prolonged, forced exposure to the CS seems, on the whole, to facilitate extinction rather than to "resensitize" the subject; and (*f*) the use of anxiety-competing responses, such as relaxation in the standard desensitization procedure, may be important in maximizing nonreinforced exposure to the aversive stimulus.

Recent years have witnessed an increasingly critical examination of theory and research in behavior modification (e.g., Franks, 1969), in marked contrast with the allegedly "defensive" polemics of earlier years (e.g., Eysenck, 1960; Eysenck & Rachman, 1965). Interest in behavior modification has been high among not only clinicians but also among experimental psychologists. The (over-) eagerness with which clinicians have embraced these "new" techniques is perhaps understandable in view of the inadequacy of those approaches and techniques that have held sway in most training and professional settings. Many experimentalists have been attracted to the field because, in our view, behavioral techniques seem to offer the opportunity to examine issues and problems of general psychological importance (viz., Lang, 1969).

Probably the technique that has attracted the most attention in both circles (though we hardly intend to suggest that the circles do not intersect—the present authors regard themselves as both clinicians and experimentalists) is Wolpe's systematic desensitization (Wolpe, 1952, 1958, 1969). As Paul (1969) has recently concluded:

> [It represents] the first time in the history of psychological treatments [that] a specific therapeutic package reliably produced measurable benefits for clients across a broad range of distressing problems in which anxiety was of fundamental importance [p. 159].

Moreover, considerable research effort has been fruitfully directed to parceling out the active ingredients of this procedure; the fact that we are still far from reliable answers is less important than the fact that much con-

---

[1] Preparation of this paper was facilitated by support from Grant MH 14557-03 from the National Institute of Mental Health. We are grateful to Marius C. Smith for his thoughtful critique of earlier drafts.

[2] Requests for reprints should be sent to G. Terence Wilson, who is now at Psychological Clinic, Rutgers University, New Brunswick, New Jersey 08903.

© 1971 by the American Psychological Association, Inc.

trolled research is underway within an experimental psychological framework.

Several reviews have appeared over the past few years examining the clinical, experimental, and theoretical foundations of desensitization (e.g., Bandura, 1969; Lang, 1969; Paul, 1969; Rachman, 1967). Aside from the inevitable fact that numerous studies have been reported since their writing, we find that none of the available reviews comes to grips with some of the very basic psychological issues raised by research in desensitization (although Lang's and Bandura's certainly are in this spirit). Especially lacking in these reviews is a careful consideration of *experimental work with infrahuman organisms* dealing with the more general problem of reducing fear and/or avoidance behavior. It seems to us peculiarly appropriate to examine these experimental bases of desensitization inasmuch as Wolpe claims to have *derived* the technique from his and others' work with animals. Moreover, many influential writers (e.g., Eysenck, 1959) argue that behavior therapy as a field is based on "learning theory," which historically has been heavily involved in the study of animal behavior. Though mindful of the inherent limitations of analogue research, we would be remiss and mistakenly "humanistic" to ignore the possible implications of research conducted, in the main, by workers who have never seen an anxiety hierarchy.

This article, then, examines experimental research in fear reduction with animals, which seems to relate fairly directly to the operations that characterize systematic desensitization, as well as to the theoretical underpinnings of the procedure. A separate paper will be devoted to a discussion of experimental research with humans.

THE PRINCIPLE OF RECIPROCAL INHIBITION

According to Wolpe, the concept of "reciprocal inhibition" is central to the technique of systematic desensitization. Originally postulated by him as a neurological mechanism to explain his early experiments with cats, it has acquired considerable surplus meaning, developing into one of behavior therapy's most frequently touted slogans. Thus, it has been referred to as a "method" (Rachman, 1967), a "learning process" (Rachman, 1965), an "explanation" (Lomont, 1965; Rachman, 1967), a "complex neural phenomenon" (Wolpe, 1962), and a "therapeutic technique" (Eysenck, 1960, 1964). Obviously such obfuscation contributes little to an understanding either of the concept or of the technique(s) that it is invoked to explain. In our opinion, it has become a panchreston, with all the unfortunate consequences attendant upon a confusion of *process* with *procedure* (cf. Evans & Wilson, 1968).

Systematic desensitization, although not a novel technique, was evolved and promulgated in its present form by Wolpe largely on the basis of his studies on the experimental induction and elimination of neurotic fears in cats. He found that analogous to the original clinical work of Jones (1924), conditioned emotional responses could be eliminated by feeding animals gradually closer to the locus of the fear conditioning. The main notion was that there is always a point along the generalization gradient at which the anxiety elicited is not so severe as to prevent the elicitation of a physiologically incompatible response, like eating (*presumed* to be incompatible)—a response that inhibits the simultaneous occurrence of the anxiety. The repeated reciprocal inhibition of the fear responses is said to result in their permanent conditioned inhibition. The conceptual framework is Hull's (1943), with "reciprocal inhibition" regarded as functionally equivalent to "reactive inhibition" (see below).

In a massive extrapolation, Wolpe substituted imaginal representation of the feared conditioned stimulus (CS) for actual exposure, and substituted Jacobson's (1938) training in muscle relaxation for feeding.

Wolpe adapted from Sherrington (1906) the term "reciprocal inhibition." Sherrington's usage referred to the suppression of one reflex by the evocation of another. Thus, reflex relaxation of the extensor muscles of the arms accompanies contraction of the flexor muscles, there being a reciprocal relaxation of the flexors when the extensors contract. As we have pointed out elsewhere (Davison, 1968), Wolpe's postulation of reciprocal inhibition at the very complex levels of neural and behavioral organization that he is concerned with represents an extreme and probably unjustified extension of the term as originally employed by Sherrington.

Wolpe explicitly rejected experimental extinction (i.e., "The specific procedure of presenting the conditioned stimulus unaccompanied by the usual reinforcement; also the decrement in a conditioned response which results from that procedure [Kimble, 1961, p. 479]") as a viable procedural description of systematic desensitization. He argued that exposure alone is ineffective as far as anxiety responses are concerned (Wolpe, 1958, p. 71), citing Miller (1951) as showing that a fear-motivated motor habit required hundreds of trials to extinguish. Aside from the fact that Miller's data demonstrated the possible *inefficiency* of repeated exposure—*not ineffectiveness*—a more detailed consideration of recent data demonstrates that conditioned avoidance responses can, indeed, be rapidly extinguished via exposure to the CS under appropriate conditions (cf. Baum, 1970; Katzev, 1967; Shearman, 1968; Solomon, Kamin, & Wynne, 1953, Experiment 5; Weiss & Monahan, 1969).

Moreover, beyond these empirical considerations, Wolpe's theoretical reasons for dismissing sheer exposure seem to be erroneous. The slow extinction of anxiety responses was attributed by him to the "fact" that autonomic responses generate only a small amount of reactive inhibition. Apart from the problems inherent in conceptualizing anxiety as solely autonomic (cf. Schachter & Singer, 1962), the basic logic of the Hullian fatigue theory of extinction—which Wolpe explicitly accepts—has been severely embarrassed (e.g., Adams, 1963; Deutsch, 1960; Gleitman, Nachmias, & Neisser, 1954; Jensen, 1961). Indeed, to return to data, Lomont (1965) has pointed out that in Wolpe's original experiments there was no control group that was gradually exposed to the locus of shock without being fed. In fact, Wolpe "cured" three cats by a "forced solution" method, entailing prevention of the avoidance response in order to forcibly expose the animals to the CS. (This datum conflicts sharply with Wolpe's dicta against exposing clients to aversive imaginal and real-life stimuli once anxiety has been elicited; cf. our discussion below of the "resensitization hypothesis.")

Thus, Wolpe's rejection of the possible effectiveness of reexposure without explicitly provided antagonistic responses might be accounted for by the following chain of three erroneous assumptions leading to an invalid conclusion: (*a*) Extinction takes place exclusively via buildup of a negative habit, *sIr*, which (*b*) requires generation of an aversive drive, *Ir*, from the response to be eliminated; (*c*) anxiety, as primarily or solely an autonomic (sympathetic) response, generates little *Ir*, so that (*d*) therefore a specific incompatible response must be elicited to work against the simultaneous evocation of the fear response, providing reciprocal inhibition, a functional equivalent of reactive inhibition, as a basis for the acquisition of the negative habit, *sIr*. The very questionable validity of Steps 1–3 renders the conclusion (*d*) invalid.

It appears, then, that Wolpe's celebrated formulation of the reciprocal inhibition principle constituted an unwarranted and probably also unparsimonious transformation of a narrow physiological principle to explain very complex and occasionally insufficiently controlled behavioral data. This certainly does not detract from the considerable practical significance of the work attendant on Wolpe's speculations, but it may serve to allow us to break set in order to examine carefully the experimental literature that is relevant to an understanding of fear reduction as it relates to desensitization.

LITERATURE REVIEW

Systematic desensitization involves at least the following crucial elements: (*a*) Exposure to aversive stimuli, (*b*) gradations in this exposure from least to most anxiety provoking, and (*c*) pairing of anxiety-competing responses with graded exposure to aversive stimuli.

Ungraded Exposure to Aversive Stimuli Paired and Unpaired with Anxiety-Competing Responses

Lomont (1965) has reviewed animal experiments in which comparisons were made in efficacy of fear reduction via repeated exposure to the CS at training intensity versus pairing of such exposures with anxiety-competing responses. Farber (1948) found that feeding in the presence of the CS eliminated conditioned fear, but lacking in the study was an appropriate control group in which animals received exactly the same exposure opportunities without the feeding. Moltz (1954) replicated Farber and added this necessary control, finding that feeding in the presence of the CS did not enhance fear reduction beyond merely confining rats in the fear-provoking situation. Sermat and Shephard (1959) found that feeding produced only a weak and transient reduction in a conditioned avoidance response (CAR) when compared with a group merely confined in the nonreinforced presence of the CS. Lane (1954) and Hall (1955) obtained results indicating that feeding and the presence of a fearless cage-mate, respectively, produced greater fear reduction than controls confined for the same time in the presence of the CS. However, in both these studies, the greater movement of the experimental animals might have produced greater reduction in fear solely in terms of increased unreinforced exposure to conditioned fear cues in the apparatus.

From all these earlier studies Lomont concluded that "reciprocal inhibition" fails to receive appreciable support as a mode of anxiety reduction superior to mere repeated exposure to the CS. However, a later study by Gale, Sturmfels, and Gale (1966), he asserted, did provide such evidence. Using a classical conditioning (conditioned emotional response—CER) rather than a CAR paradigm, these workers found that providing "food as a stimulus for a response incompatible with fear" reduced fear more quickly than mere exposure to the same graded CS series. The CER paradigm avoided the problems of increased exposure due to greater movement in one group versus the other, but the use of defecation as the dependent variable when food is given to one group and not to the other raises problems of interpretation. This could be handled by a control in which food is provided *unpaired* with anxiety-eliciting stimuli. Of more fundamental importance perhaps is the fact that they did not include a regular extinction control group in which the CS was presented at its original intensity. Without this group it is impossible to conclude that "extinction" is less effective than "reciprocal inhibition." Indeed, Poppen (1968), as discussed below, has shown that a hierarchy does not produce quicker extinction than regular CS presentation within the CER paradigm. Since Gale et al. obtained no significant differences between their experimental groups on posttreatment *test trials,* we cannot conclude that either of their groups is more *effective* than regular extinction. In contrast to Lomont's advised caution, Rachman's (1967) statement that this particular study "clearly indicates the superiority of the inhibition technique [p. 98]" seems unjustified.

Recent experiments by Nelson (1966, 1967) contain controls that shed new light on the possible role of stimuli that supposedly elicit anxiety-competing responses. In the first study (Nelson, 1966) rats were conditioned to a CS, following Lane (1954), and then were confined in the fear compartment of an apparatus like that introduced by Miller (1948), with or without food (when hungry), and were finally tested for degree of residual fear. Escape from the fear compartment was prevented during conditioning and allowed only during the last sequence. Level of fear was thereby inferred by evaluating its strength as a drive for learning a new response. This represented an improvement on Lane's design, where animals were given shock-escape training in the first phase of the experiment, the dependent measure being latency of subsequent escapes from the fear compartment. Lane had found that rats confined without food crouched at the escape door while the animals confined with food present moved around the compartment. Nelson placed the rats directly over the food dish, thereby minimizing the conditioning of approach responses to food that may have confounded the measurement of fear, as in Lane's experiment. Results indicated that confinement per se, with or without food, reduced fear significantly more than no confinement. Comparisons of food versus confinement groups showed the *former* to be more fearful. In contrast to Lane's findings, Nelson reported that feeding resulted in *less* exposure to the fear cues in the shock compartment, since his procedure of placing hungry animals directly over the food dish resulted in keeping them in the area of the food dish rather than roaming around the compartment.

In a second experiment, Nelson (1966) added the following interpolated treatments in order to isolate better the relationship between exposure to the CS and the presence of food in the shock box: (*a*) Incentive: food present in shock compartment but freedom to move from one side to the other; (*b*) no incentive: same freedom but without food; (*c*) incentive control: confinement in the shock box to match the time spent by a yoked rat in the incentive group, except with no food present; and (*d*) no incentive control: similarly matched with the no incentive group, with no food present. Results reveal that the presence of food facilitated fear reduction *by serving as an incentive for exposure.*

In the same vein, Nelson (1967) ran a study employing chlorpromazine as a means to compare sheer exposure with exposure facilitated by competing responses. Rats were conditioned undrugged, given exposure to the fear compartment with or without the drug, and then tested undrugged for residual fear. He found that drugged rats were less fearful than undrugged animals only when the drug resulted in increased exposure to the fear compartment.

Taken together, Nelson's three studies indicate strongly that the provision of a stimulus that supposedly elicits an anxiety-competing response (food, chlorpromazine) facilitates extinction of fear through the increased exposure to the CS which it can bring about. When these additional stimulus conditions are arranged so as *not* to provide such increased exposure, as in his first study, they result in *poorer* reduction of fear. Thus, by focusing less on the competing responses and more on the stimulus conditions that supposedly effect these responses, the importance of feeding (as stressed by Wolpe) is seen in a different light. Indeed, Wolpe (1958) himself reported that anxiety in his cats was more rapidly eliminated when, in feeding them in the shock compartment, "the pellets were tossed at widely distributed points than if they were confined to the food box [p. 57]."

Wilson and Dinsmoor (1970) have concluded that feeding per se can reduce fear. They established passive avoidance behavior in rats by shocking them for leaving a small platform at the end of a grid runway. During extinction treatment trials, one group was fed on the platform, a regular extinction group was simply placed on the platform, and a third group was forcibly exposed to the grid. On test trials both the fed and forced exposure groups showed significantly greater extinction than the regular extinction group, but apparently did not differ from each other. Wilson and Dinsmoor reported that feeding the rats tended to decrease rather than increase exposure to fear stimuli, and they pointed out that neither the reinforcement of incompatible motor responses nor freezing can explain the results, since these activities would have kept the rats on the platform and so delayed rather than facilitated extinction in the passive avoidance paradigm.

A dissertation by Heath (1969) purports to furnish evidence in favor of the overriding importance of sheer exposure as compared to "reciprocal inhibition," but several methodological problems make his conclusions doubtful. In addition to a "disuse" control group, he compared the efficacy of two special treatment procedures in the extinction of avoidance behavior in a Mowrer-Miller shuttle box, tonal CS duration during acquisition being 10 seconds: "Reciprocal inhibition" rats were provided food during five 100-second exposures to a preprogrammed five-step tonal hierarchy, with no escape possible during the extended presentations of the generalized CS, and "flooding" rats had five 100-second exposures to the CS at training intensity without food available. Since both these groups were superior to the no-treatment ("disuse") group and not different from each other, Heath concluded that exposure alone to the CS was the critical factor, with "reciprocal inhibition" playing no role. The problems we see are (*a*) the presence of the food tray, from which reciprocal inhibition animals could eat, introduced a difference in stimulus conditions during treatment as compared to the flooding animals; (*b*) reciprocal inhibition subjects had extended and forced exposure to the tonal CS, making this an inappropriate analogue of

Wolpe's procedure; and (*c*) the absence of a regular extinction condition (repeated exposure to the 10-second CS at training intensity) precludes any judgment on the relative efficacy of the special treatment procedures. There also are serious questions as to whether the reciprocal inhibition procedures were such as to allow feeding to take place at all during treatment; this more general problem of pairing is discussed below.

Gambrill (1967) employed a rather complex design in an attempt to compare the relative effectiveness of reinforcing a response incompatible with a previously acquired avoidance response on the elimination of the latter. After rats were conditioned to reliably avoid a shock on a Sidman schedule by turning a wheel, certain groups were given training in lifting a lever for food. When compared with a traditional extinction group, superior extinction of one of the "counterconditioning" groups was manifest only on the first test session. Afterward, for the remaining nine extinction test sessions, no differences were observed among the various groups. Her conclusion was that providing an instrumental appetitive response incompatible with the avoidance response was not superior to only nonreinforcement in reducing avoidance.

Counterconditioning and the problem of pairing. Gambrill's discussion of these results leads to some comments we feel appropriate to make at this point regarding the adequacy of the designs in the foregoing animal studies to test the effects of counterconditioning. We find all these studies deficient in how they operationalize "counterconditioning" in that no provision is made for the aversive stimuli to be removed as soon as the incompatible response does not dominate over the anxiety/avoidance. One must keep in mind the standard desensitization procedure, in which the individual is not only presented with a graded series of anxiety-provoking stimuli (as, indeed, most of the above studies do *not* do), at the same time given the opportunity to perform an incompatible response, but most importantly, is enabled—nay, instructed—to remove himself from the aversive situation as soon as the anxiety occurs—that is, as soon as the response considered antagonistic to that to be counterconditioned ceases to be dominant. It is only in this way that one ensures that, as much as possible, the organism is confronting the original aversive situation (initially a watered-down form of it) without behaving fearfully. Thus, careful programming of the aversive stimuli is necessary in order to minimize the occasions on which the organism is performing in the presence of the CS that response which is the target for removal; providing only for graded presentation of the CS is probably necessary, but rarely sufficient as well to ensure true pairing. Of course, if one is concerned only with pairing at the *onset* of the CS, then studies such as those reviewed above probably satisfy the criterion; that is, the animal eats as the CS *comes on*. However, as often happens in desensitization, the CS at some point *disrupts* the competing activity. It is this disruption that we are talking about; to the degree that the anxiety-competing response does not persist *throughout* CS presentation we do not see pairing taking place. We should point out also that our conceptualization of counterconditioning is more delimiting than Kimble's usage, namely, "Extinction under circumstances in which the response decrement is hastened by reinforcement of a response which displaces the original conditioned response [Kimble, 1961, p. 478]."

Graded Exposure to Aversive Stimuli Paired and Unpaired with Anxiety-Competing Responses

Systematic desensitization involves not only a pairing of competing responses with aversive stimuli but also the use of a graded hierarchy according to which these stimuli are presented to the person. The argument has been that the use of such a hierarchy makes it more likely that anxiety-competing relaxation will predominate over the anxiety ordinarily elicited by the particular generalized aversive stimulus. We have seen from the section just above that considerable reduction of anxiety/avoidance is possible when animals are exposed repeatedly to the full-intensity CS, and that, furthermore, this reduction is usually not inferior to that brought about when the animal is perhaps enabled to perform an incompatible response during CS presentations. The only studies thus far covered that employed any kind of graduated presentations were those by Gale et al. (1966) and Heath (1969), though the problems already pointed out make it impossible to evaluate the necessity or desirability for graded presentations. Several studies have, however, been executed with animals which, in an analogue fashion, would seem to have bearing on the use of an anxiety hierarchy in desensitization in conjunction with anxiety-competing responses.

In the review already cited, Lomont (1965) concluded that the bulk of experimental data with rats did not indicate that progressive exposure to the CS results in faster extinction than exposure to it at training intensity. Most of the studies discussed were within the Murray-Berkun (1955) displacement model, in terms of which, following the establishment of an approach-avoidance conflict, rats which have been allowed to make goal responses in "displaced" alleys (alleys dissimilar to the original one in width and color) eventually make approach responses in the original alley. While confirming this, Berkun (1957) did show that when total number of trials were considered, rats given displacement training required slightly *more* trials to reach a criterion of recovery than did control animals given trials only in the original alley. This negative finding was replicated by Noblin and Maher (1962) and Taylor and Rennie (1961). On the other hand, Taylor and Maher (1959) did find an economic advantage from the displacement trials. In the absence of any definitive account of the complex variables involved in displacement responding, however, Noblin and Maher's (1962) caution in interpreting these several studies seems appropriate. At any rate, these displacement studies are an inexact analogy to desensitization, in which gradual progression involves a typically lengthy hierarchy of situations, not just one or two slightly differing ones.

A more direct test of the importance of gradual stimulus presentation was attempted in Kimble and Kendall's (1953) comparison of Guthrie's (1935) "toleration" method of producing extinction with the conventional method, which they term "exhaustion." They found that the former was significantly more effective, and concluded therefrom that overt nonreinforced avoidance responses were not as crucially necessary for extinction as Hull's inhibitory postulates had held. Weiss and Monahan (1969) criticized this widely cited study, pointing out that Kimble and Kendall's toleration rats had considerably more exposure to the CS or variations thereof than did exhaustion animals. This disparity makes it impossible to conclude anything about *progressive* exposure to the CS. Additionally, one can pose questions about the nature of their "toleration" condition. Subjects were given seven "toleration trials" in *each of which* the CS intensity was gradually raised from some initially lower-than-training level to the full training level. It is important to note that this procedure differs from one in which successive levels of the original CS are presented as discrete trials, and where progress to a higher level of CS intensity is *contingent* on the cessation of responses at the immediately preceding level. Such a response-contingent hierarchy we define as toleration, and we distinguish it from the presentation of a preprogrammed series of graded stimuli where progress up the hierarchy is *noncontingent* on the absence of an avoidance response (cf. Gale et al., 1966; Heath, 1969).

In a tightly controlled experiment designed to overcome the methodological problems of the Kimble-Kendall study, Weiss and Monahan (1969) matched rats with respect to a stringent criterion of avoidance proficiency. Toleration subjects, during the "treatment" phase, progressed through a hierarchy of the visual training CS, beginning at 25 volts and increasing by 5 volts on each trial, provided no avoidance response was made (thus fulfilling the criterion we have suggested above vis-à-vis analogue studies of the clinical desensitization procedure). Exhaustion subjects were *yoked* to toleration subjects, in the same fashion employed by Davison (1968), that is, number and duration of exposures to full-intensity CS being determined by the behavior of the toleration "partner." Results showed a striking superiority of the toleration group in extinguishing the avoidance response.

In addition to differences in the dependent measure—number of avoidance responses on test extinction trials—toleration subjects also emitted fewer avoidance responses during their treatment, in fact, about one-fourth of those emitted by their yoked exhaustion mates. In order to achieve precise control over sheer exposure to the CS and its variations (the problem noted with the Kimble-Kendall study), the warning light remained on during all treatment trials for both groups, in spite of whether an avoidance response occurred. The success of the toleration procedure probably lay in significantly reducing the number of avoidance responses during treatment. This served both to expose those animals to more unreinforced presentations of the training CS (albeit via its generalization gradient), which is known to facilitate extinction (Kalish, 1954), and to reduce response-maintaining "safety signals" associated with the occurrence of an avoidance response (Bolles & Grossen, 1969). It should be noted that Weiss and Monahan employed a nonconventional procedure in which the occurrence of an avoidance response did not terminate the CS, which remained on throughout each trial. This and other methods of forced exposure to the CS are discussed below.

Thus far there have been only two studies attempting, within a single design, to compare the efficacy of graded versus ungraded exposure to the CS paired and unpaired with anxiety-inhibiting stimuli. Goldstein (1969) attempted to evaluate the advantages of providing an incompatible response in addition to arranging for progressive exposure to a feared stimulus in an approach-avoidance situation. While he concluded that the anxiety-competing response (eating) facilitated anxiety reduction, a careful examination of his statistical analysis fails to justify the conclusion: not only did an analysis of variance fail to yield a significant F ratio, but none of the treatment groups did, in fact, differ significantly from each other.

Poppen (1968) chose the classical conditioning paradigm of conditioned suppression as an analogue to systematic desensitization, rejecting avoidance conditioning on the grounds that it is a problematical measure of fear (see below). After teaching rats to bar press very reliably for food and then to suppress such pressing during a tonal CS, he assigned animals to one of several conditions, carefully matching for acquisition of the conditioned suppression: (*a*) "Regular extinction," whereby the CS was always presented at original intensity; (*b*) "inhibition," in which, during the CS period, food was provided gratis (this presumably being, then, ungraded exposure paired with food); (*c*) "toleration," in which subjects progressed through five generalized tones before reaching training frequency, progress being contingent on a criterion of extinction at each tonal frequency (similar, therefore, to the toleration group of Weiss and Monahan, 1969); and (*d*) "toleration plus inhibition," in which free pellets were made available during the same kind of graded presentation as in the toleration group (hence an analogue to desensitization).

Two principal dependent measures were used, and the results differ occasionally depending on which measure is examined. In terms of mean number of trials necessary to reach criterion of extinction, all groups but Toleration + Inhibition did not differ significantly, the latter alone being significantly superior to *all* the others. The striking superiority of this analogue-to-desensitization group seemed to account for the significant effects associated with the two factors, type of CS (graded versus full) and presence or absence of food. Providing a hierarchy did not prove superior to sheer repeated presentations of the original CS, and providing food paired with the CS expedited extinction only when the CS was presented gradually; associating a presumably anxiety-competing response with repeated presentations of the CS at training intensity did not expedite extinction over and above sheer repeated exposure, results consonant with Moltz (1954) and Nelson (1966).

Whatever the merit of Poppen's reasoning about an adequate measure of fear, however, a classical conditioning paradigm, unlike the avoidance situation, does not meet a fundamental requirement of systematic desensitization, namely, that exposure to the aversive stimulus be under the control of the subject and not the experimenter. Although the rats were said to have eaten most of their pellets during treatment sessions, which suggests that fear was not such as would have led to much escaping from the CS, the procedural discrepancy with systematic desensitization must be considered.

Additional findings can be mentioned. During each "treatment" session required to achieve the criterion of extinction, groups did differ with respect to the amount of response suppression ("anxiety") that occurred. It is logically possible for two rats to take the same number of sessions in order to extinguish and yet to "suffer" differentially while doing so. Poppen employed a "suppression ratio" in order to make this comparison, and, when comparing the groups on this measure, some interesting differences did emerge. An analysis of variance once again showed effects for mode of CS presentation (graded versus full) and availability of an antagonistic response (food versus no food), but now many between-group differences emerged. Once again Toleration + Inhibition surpassed the other three groups in terms of extinction, or rather ease of extinction. Regular extinction was shown to be significantly inferior to all other groups, which means that either providing food (inhibition) or a graded hierarchy (toleration) reduces the amount of fear during treatment.

Thus, one can say that considerable support is found for the superiority of graded exposure being paired with anxiety-competing responses as compared to graded exposure alone, and some support is given for the superiority of graded exposure over repeated exposure to the CS at training intensity. Moreover, with regard to the issue of repeated exposure to the CS paired or unpaired with anxiety-antagonistic responses, some support is given for the preference of the former over the latter.

In summary, it is clear that all too few methodologically sound studies have been executed to test the efficacy of a progressive hierarchy, with or without accompanying anxiety-competing responses. From the above, it seems wise to conclude tentatively that gradual exposure to the CS might be preferable to the same amount of exposure to the original CS, at least when the avoidance response is not prevented (Weiss & Monahan, 1969), particularly if one is concerned that the process of treatment be as unstressful as possible (Poppen, 1968; Weiss & Monahan, 1969). The only experiment adequately designed to assess the gains from adding an anxiety-competing response to the graded exposure (Poppen) confirms the desirability of providing such an anxiety-inhibiting response, with respect both to ease and efficiency of extinction treatment.

"Resensitization": The Issue of Forced (Prolonged) Exposure

A frequently stressed procedural prescription in systematic desensitization is that as little anxiety as possible be elicited in the subject/client. This goal is pursued through (*a*) the imposition of anxiety-competing relaxation, (*b*) the use of a carefully calibrated anxiety-hierarchy, and (*c*) the provision of escaping from an aversive scene as soon as it evokes even the slightest degree of perceived discomfort. According to Wolpe, not only will the autonomic effects of relaxation be able to "reciprocally inhibit" only relatively weak anxiety responses, but "Exposure, and prolonged exposure in particular, to a very disturbing scene can seriously increase phobic sensitivity [Wolpe, 1969, p. 127]." We have already seen, however, that this integral part of the clinical lore in behavior therapy is contradicted by the well-established phenomenon of experimental extinction. And as we have stated elsewhere (Davison, 1968), the experimental literature lacks not only data showing acquisition or reinforcement of stable conditioned fear responses in humans via pairing with supposedly secondary reinforcers(like an anxiety image), but is notably lacking as well in demonstrations of stable conditioned fear in humans through reinforcement via painful electric shock as the unconditioned stimulus (UCS).[3]

---

[3] Wolpe has often cited a study by Campbell, Sanderson, and Laverty (1964), in which alcoholics heard a tonal CS during 2 minutes of muscular paralysis induced by an injection of Scoline, which pairing resulted in an apparently conditioned fear reaction extremely resistant to extinction; rather than viewing this experiment as demonstrating a typical aspect of human neurotic behavior, we see it as a bizarre demonstration that bears only the most remote similarity to what is found in civilian life, certainly far removed from the kind of forced exposure to a distressing image that Wolpe cautions

On the other hand, there are indeed scattered research findings, intriguing and paradoxical, that may lend some plausibility to Wolpe's clinical assertion regarding resensitization, and these should be mentioned as well. Eysenck (1968) has cited an experiment by Napolkov in which unreinforced presentation of the CS after only one CS-UCS pairing produced *increasing* levels of blood pressure. In fact, the conditioned response (CR) rapidly grew greater than the original unconditioned response (UCR)! More recently, Rohrbaugh and Riccio (1970) have demonstrated what they term the "paradoxical enhancement" of fear in rats. In one experiment rats were exposed to unreinforced conditioned fear stimuli for 0, ½, 5, 15, or 50 minutes. Fear was subsequently assessed in terms of suppression of approach to food or water with the CS present. While the 15- and 50-minute exposures resulted in some extinction (the usual finding), the 5-minute group showed *more* residual fear than nonexposed controls. In a second experiment, 30- and 60-second interpolated exposures to the CS increased sensitivity as compared to 300-second exposure subjects and nonexposed controls. The investigators concluded that paradoxical enhancement is likely to be an exceptional rather than a typical outcome of unreinforced exposure to the CS, though the necessary and sufficient conditions for its occurrence must surely be delineated through careful parametric studies.

Forced exposure to an aversive CS. Most of the avoidance studies discussed in this paper are conceptualized in terms of two-factor theory of avoidance learning (Rescorla & Solomon, 1967). This view emphasizes the *classical extinction* of fear, which is said to motivate the instrumental response. This is optimally achieved by nonreinforced, inescapable exposure to the CS which prevents any short-latency avoidance response that might *conserve anxiety* by removing the subject from the CS before the fear CR develops (Solomon & Wynne, 1954).

Baum (1970) has systematically explored the effect of preventing an avoidance response on extinction. He has found that the *duration* of the response prevention used is a critical factor influencing the results achieved with this method. In general, the greater the duration of forced exposure to the CS, the greater the extinction; too short an exposure has little effect on extinction. The more resistant to extinction the response is (viz., the greater the "anxiety" is as a result of a high intensity US or increased amounts of shock exposure), the longer is the period of response prevention necessary to eliminate the avoidance response. The extinction-producing effect of response prevention can be significantly facilitated by either introducing nonfearful rats into the apparatus (such as Hall, 1955, did) or forcing the rat to move around the experimental compartment. Systematic observation of the rats' behavior during response prevention and the success of "social and mechanical facilitation" strategies suggest that eliciting nonfearful behavior during exposure to the aversive CS is an important variable in producing extinction of an avoidance response.

Page (1955) had earlier reported that response prevention or blocking facilitated extinction of the avoidance response not because fear had been reduced, but because confinement resulted in subjects learning a "freezing" response which was incompatible with the measurement of avoidance. An independent measure of approaching the locus of shock for food revealed a residual fear in subjects whose responses had been blocked. Coulter, Riccio, and Page (1969) have replicated these results and found greater residual fear in rats undergoing response prevention than in rats extinguished in the regular manner. Kamin, Brimer, and Black (1963) found that a CS, to which the avoidance response had been extinguished, retained aversive properties for subjects as demonstrated in a conditioned suppression paradigm. Evidence of this nature has led Poppen (1968) and Bandura (1969)

---

against. Moreover, a subsequent experiment by the same research group (Madill, Campbell, Laverty, Sanderson, & Vanderwater, 1966) failed to show differences in autonomic reactions and avoidance behavior following true conditioning (CS paired with UCS) as compared with pseudo-conditioning (UCS alone). In addition, a more recent paper by Farrar, Powell, and Martin (1968) reported that following such treatment with 12 alcoholics, the 9 patients located at a 1-year follow-up showed little effect of the traumatic conditioning treatment.

to reject the extinction of avoidance conditioning by means of blocking as an appropriate analogue of fear reduction in systematic desensitization. However, Nelson (1969) has shown that blocking can have a significant fear-reducing effect. In contrast to his and other studies cited above, Nelson conditioned fear, not by escape training, but by punishing entry into the shock compartment of a Mowrer-Miller box. As in the Coulter et al. (1969) and Page (1955) studies, fear was assessed by entry into the shock compartment for food, a procedure which did not confound freezing with fear extinction. Similarly, using a passive avoidance paradigm (as discussed above), which rules out an explanation in terms of the reinforcement of incompatible motor responses or freezing, Wilson and Dinsmoor (1970) found that blocking produced more rapid extinction than the regular extinction method.

It is important to note that another major line of theorizing and research which eschews two-factor theory in explaining avoidance learning also prescribes response prevention as an optimal extinction procedure. The *discrimination theory* holds that extinction is expedited to the degree that the subject can distinguish the change in reinforcement contingencies between acquisition training and extinction (Bolles & Grossen, 1969; D'Amato, 1970; Herrnstein, 1969). Blocking eliminates the *response-contingent termination of the CS* which is characteristic of the regular extinction procedure, and which precludes a discrimination of altered reinforcement contingencies. This critical discrimination is also facilitated by "flooding," or the procedure in which the CS remains on for an extended time, irrespective of the subject's response, and which has been shown to be superior to regular extinction (Katzev, 1967; Polin, 1959; Shearman, 1968).

The procedural implications for desensitization are clear: Wolpe prescribes that the client/subject be able to escape from a scene as soon as it becomes aversive, and if one construes the anxiety signal as an avoidance response, then desensitization is unnecessarily inefficient. It is interesting to note that Lomont (1965), in arguing for the merits of the "extinction" interpretation of systematic desensitization, advanced the "fact" that the operations of systematic desensitization were optimal for producing extinction insofar as the client is not allowed to escape from the aversive stimuli. In fact, it is the client, and not the therapist, who determines the length of exposure to the aversive scene in systematic desensitization.[4]

CONCLUSION

The foregoing review should make it clear that easy generalizations about the reduction of human fear with limited reference to animal studies are misleading. The complex, incomplete, and often inconsistent nature of the extinction literature that Kimble (1961) noted is still evident, especially that area dealing with the elimination of fear/avoidance behavior. Nevertheless, some general conclusions can be drawn. The counterconditioning procedure has been shown to be more effective and less disruptive in reducing classically conditioned fear than the orthodox extinction procedures using conditioned suppression and passive avoidance paradigms (Poppen, 1968; Wilson & Dinsmoor, 1970). With avoidance behavior the provision of counterconditioning agents like food and chlorpromazine probably expedites extinction by facilitating exposure to the CS which would otherwise be avoided (Nelson, 1967). Weiss and Monahan (1969) have demonstrated how the toleration procedure achieves the same end.

Extrapolating from these laboratory findings to the clinical setting, the provision of anxiety-competing relaxation with graded exposure to aversive stimuli may be necessary with clients for reasons other than those originally held. A conceptualization of desensitization in terms of the extinction rather than counterconditioning paradigm is not inconsistent with the necessity for relaxation. For it may be that rather than providing for "reciprocal inhibition," since anxiety generates too little "reactive inhibition" to form the basis of "conditioned inhibition," relaxation functions as did food and chlorpromazine in Nelson's studies (1966, 1967), namely, as a *facilitator for exposure.* It must be remembered that phobics are, by definition, people who typically avoid their respective feared objects. In order to encourage them to expose themselves to what they fear, it may be necessary not only to create a situation that is graduated, beginning with relatively mild "doses" of the phobia, but also to reduce still further the distress to be elicited by stimuli far along the generalization gradient by providing the person with a comforting antianxiety response, like relaxation. Since progress up to the hierarchy is traditionally controlled by the client, he should be more likely to confront aspects of his fear if he is bolstered in some way, for example, by knowing how to relax. Morever, a recent experiment by Geer, Davison, and Gatchel (1970) with human subjects shows that electric shocks cause significantly less distress if subjects only *believe* they have control over their duration, even when that control is illusory.

---

[4] The problems with analogues are well-illustrated by the issue of the anxiety signal as an avoidance response. At first blush, it seems rather straightforward to assume that when an individual undergoing desensitization signals anxiety and thereby is removed from the aversive image, he is performing a response functionally identical to, for example, a wheel-turning response by a rat in a typical avoidance apparatus. However, it is customary for a human subject to know beforehand what the procedure of desensitization entails, namely, graduated and *repeated* exposure to imaginal aversive stimuli. Thus, when the human subject "avoids" (or better, "escapes"), is this really the same as the wheel turn? We doubt it, for the human subject knows that *because* of this anxiety signal, he will have to confront the same stimulus again.

The critical role of sheer exposure to the CS is highlighted by studies of avoidance behavior which demonstrate that response prevention is an effective and efficient method of extinction. These studies have led to therapeutic strategies relying on extended forced exposure to the CS (Stampfl, 1966). These techniques construe the client's anxiety signal in desensitization as an avoidance response and are directed toward preventing client-controlled termination of the aversive image in the belief that such escaping will simply serve to maintain the fear.

Systematic desensitization has been predicated by Wolpe and his colleagues on the allegedly sound experimental basis established in the learning laboratory. This review provides support for some of the *procedural* prescriptions of desensitization at the same time that it raises serious questions about the conditioning explanations put forth by Wolpe. Particular issue is taken with the concept of "reciprocal inhibition," not the least of which is that it represents a shaky neurologization of a behavioral system (Hull) which has already been shown to have serious logical and empirical shortcomings. In addition, experimental findings point to possibly new therapeutic techniques other than those which Wolpe originally allowed. We would now anticipate, in addition to the sorely needed animal research, systematic investigation into the processes of the technique itself, introducing modifications that are truly derived from more basic research. Findings from such research efforts will help pin down the analogue results and contribute both to the improvement of the (changed) technique and to the clarification of its process-mechanism underpinnings.

REFERENCES

ADAMS, J. A. Comment on Feldman's "Reconsideration of the extinction hypothesis of warm up in motor behavior." *Psychological Bulletin,* 1963, **60**, 460–463.

BANDURA, A. *Principles of behavior modification.* New York: Holt, 1969.

BAUM, M. Extinction of avoidance responding through response prevention (flooding). *Psychological Bulletin,* 1970, **74**, 276–284.

BERKUN, M. M. Factors in the recovery from approach-avoidance conflict. *Journal of Experimental Psychology,* 1957, **54**, 65–73.

BOLLES, R. C., & GROSSEN, N. E. Effects of an informational stimulus on the acquisition of avoidance behavior in rats. *Journal of Comparative and Physiological Psychology,* 1969, **68**, 90–99.

CAMPBELL, D., SANDERSON, R. E., & LAVERTY, S. G. Characteristics of a conditioned response in human subjects during extinction trials following a single traumatic conditioning trial. *Journal of Abnormal and Social Psychology,* 1964, **68**, 627–639.

COULTER, X., RICCIO, D. C., & PAGE, H. A. Effects of blocking an instrumental avoidance response: Facilitated extinction but persistence of "fear." *Journal of Comparative and Physiological Psychology,* 1969, **68**, 377–381.

D'AMATO, M. R. *Experimental psychology.* New York: McGraw-Hill, 1970.

DAVISON, G. C. Systematic desensitization as a counterconditioning process. *Journal of Abnormal Psychology,* 1968, **73**, 91–99.

DEUTSCH, J. A. *The structural basis of behavior.* Chicago: University of Chicago Press, 1960.

Evans, I., & Wilson, G. T. Note on the terminological confusion surrounding systematic desensitization. *Psychological Reports,* 1968, 22, 187–191.

Eysenck, H. J. Learning theory and behavior therapy. *Journal of Mental Science,* 1959, **105**, 61–75.

Eysenck, H. J. (Ed.) *Behaviour therapy and the neuroses.* New York: Pergamon Press, 1960.

Eysenck, H. J. (Ed.) *Experiments in behaviour therapy.* New York: Macmillan, 1964.

Eysenck, H. J., & Rachman, S. J. *The causes and cures of neurosis.* San Diego: Knapp, 1965.

Farber, I. E. Response fixation under anxiety and non-anxiety conditions. *Journal of Experimental Psychology,* 1948, **38**, 111–131.

Farrar, C. H., Powell, B. J., & Martin, L. K. Punishment of alcohol consumption by apneic paralysis. *Behaviour Research and Therapy,* 1968, **6**, 13–16.

Franks, C. M. (Ed.) *Behavior therapy: Appraisal and status.* New York: McGraw-Hill, 1969.

Gale, D. S., Sturmfels, G., & Gale, E. N. A comparison of reciprocal inhibition and experimental extinction in the therapeutic process. *Behaviour Research and Therapy,* 1966, **4**, 149–156.

Gambrill, E. Effectiveness of the counterconditioning procedure in eliminating avoidance behavior. *Behaviour Research and Therapy,* 1967, **5**, 263–274.

Geer, J. H., Davison, G. C., & Gatchel, R. Reduction of stress in humans through nonveridical perceived control of aversive stimulation. *Journal of Personality and Social Psychology,* 1970, **16**, 731–738.

Gleitman, H., Nachmias, J., & Neisser, U. The S-R reinforcement theory of extinction. *Psychological Review,* 1954, **61**, 23–33.

Goldstein, A. J. Separate effects of extinction, counterconditioning and progressive approach in overcoming fear. *Behaviour Research and Therapy,* 1969, **7**, 47–56.

Guthrie, E. R. *The psychology of learning.* New York: Harper, 1935.

Hall, J. C. Some conditions of anxiety extinction. *Journal of Abnormal and Social Psychology,* 1955, **51**, 126–132.

Heath, G. H. A comparison of reciprocal inhibition and flooding in decreasing the strength of a conditioned avoidance response. *Dissertation Abstracts,* 1969, **29**, 1842B–1843B.

Herrnstein, R. J. Method and theory in the study of avoidance. *Psychological Review,* 1969, **76**, 49–69.

Hull, C. L. *Principles of behavior.* New York: Appleton-Century-Crofts, 1943.

Jacobson, E. *Progressive relaxation.* Chicago: University of Chicago Press, 1938.

Jensen, A. R. On the reformulation of inhibition in Hull's system. *Psychological Bulletin,* 1961, **58**, 274–298.

Jones, M. C. A laboratory study of fear: The case of Peter. *Pedagogical Seminary,* 1924, **31**, 308–315.

Kalish, H. I. Strength of fear as a function of the number of acquisition and extinction trials. *Journal of Experimental Psychology,* 1954, **47**, 1–9.

Kamin, L. J., Brimer, C. J., & Black, A. H. Conditioned suppression as a monitor of fear of the CS in the course of avoidance training. *Journal of Comparative and Physiological Psychology,* 1963, **56**, 497–501.

Katzev, R. Extinguishing avoidance responses as a function of delayed warning-signal termination. *Journal of Experimental Psychology,* 1967, **75**, 339–344.

Kimble, G. A. *Hilgard and Marquis' conditioning and learning.* New York: Appleton-Century-Crofts, 1961.

Kimble, G. A., & Kendall, J. W. A comparison of two methods of producing experimental extinction. *Journal of Experimental Psychology,* 1953, **45**, 87–90.

Lane, B. R. The reduction of anxiety under three experimental conditions. Unpublished doctoral dissertation, Columbia University, 1954.

Lang, P. J. The mechanics of desensitization and the laboratory study of fear. In C. M. Franks (Ed.), *Behavior therapy: Appraisal and status.* New York: McGraw-Hill, 1969.

Lomont, J. F. Reciprocal inhibition or extinction? *Behaviour Research and Therapy,* 1965, **3**, 209–219.

Madill, M., Campbell, D., Laverty, S. G., Sanderson, R. E., & Vanderwater, S. L. Aversion treatment of alcoholics by succinylcholine-induced apneic paralysis. *Quarterly Journal of Studies on Alcoholism,* 1966, **27**, 483–509.

Miller, N. E. Studies of fear as an acquirable drive: I. Fear as motivation and fear-reduction as reinforcement in the learning of new responses. *Journal of Experimental Psychology,* 1948, **38**, 89–101.

Miller, N. E. Learnable drives and rewards. In S. S. Stevens (Ed.), *Handbook of experimental psychology.* New York: Wiley, 1951.

Moltz, H. Resistance to extinction as a function of variations in stimuli associated with shock. *Journal of Experimental Psychology,* 1954, **47**, 418–424.

Murray, E. J., & Berkun, M. M. Displacement as a function of conflict. *Journal of Abnormal and Social Psychology,* 1955, **51**, 47–56.

Nelson, F. Effects of two counterconditioning procedures on the extinction of fear. *Journal of Comparative and Physiological Psychology,* 1966, **62**, 208–213.

Nelson, F. Effects of chlorpromazine on fear extinction. *Journal of Comparative and Physiological Psychology,* 1967, **64**, 496–498.

Nelson, F. Free versus forced exposure with two procedures for conditioning and measuring fear in the rat. *Journal of Comparative and Physiological Psychology,* 1969, **69**, 682–687.

Noblin, C. D., & Maher, B. T. Temporal and physical factors in avoidance reduction. *Journal of Comparative and Physiological Psychology,* 1962, **55**, 62–65.

Page, H. A. The facilitation of experimental extinction by response prevention as a function of the acquisition of a new response. *Journal of Comparative and Physiological Psychology,* 1955, **48**, 14–16.

PAUL, G. L. Outcome of systematic desensitization. II. Controlled investigations of individual treatment, technique variations, and current status. In C. M. Franks (Ed.), *Behavior therapy: Appraisal and status*. New York: McGraw-Hill, 1969.

POLIN, A. T. The effect of flooding and physical suppression as extinction techniques on an anxiety-motivated avoidance locomotor response. *Journal of Psychology,* 1959, **47**, 253–255.

POPPEN, R. L. Counterconditioning of conditioned suppression. Unpublished doctoral dissertation, Stanford University, 1968.

RACHMAN, S. Studies in desensitization—I: The separate effects of relaxation and desensitization. *Behaviour Research and Therapy,* 1965, **3**, 245–252.

RACHMAN, S. Systematic desensitization. *Psychological Bulletin,* 1967, **67**, 93–103.

RESCORLA, R. A., & SOLOMON, R. L. Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychological Review,* 1967, **74**, 151–182.

ROHRBAUGH, M., & RICCIO, D. C. Paradoxical enhancement of learned fear. *Journal of Abnormal Psychology,* 1970, **75**, 210–216.

SCHACHTER, S., & SINGER, J. E. Cognitive, social, and physiological determinants of emotional state. *Psychological Review,* 1962, **69**, 379–399.

SERMAT, V., & SHEPHARD, A. H. The effect of a feeding procedure on persistent avoidance responses in rats. *Journal of Comparative and Physiological Psychology,* 1959, **52**, 206–211.

SHEARMAN, R. W. Response-contingent CS termination in extinction of avoidance learning. Unpublished doctoral dissertation, University of California at Los Angeles, 1968.

SHERRINGTON, C. S. *The integrative action of the nervous system.* New Haven: Yale University Press, 1906.

SOLOMON, R. L., KAMIN, L. J., & WYNNE, L. C. Traumatic avoidance learning: The outcomes of several extinction procedures with dogs. *Journal of Abnormal and Social Psychology,* 1953, **48**, 291–302.

SOLOMON, R. L., & WYNNE, L. C. Traumatic avoidance learning: The principles of anxiety conservation and partial irreversibility. *Psychological Review,* 1954, **61**, 353–385.

STAMPFL, T. G. Implosive therapy: The theory, the subhuman analogue, the strategy, and the technique. In S. G. Armitage (Ed.), *Behavior modification techniques in the treatment of emotional disorders.* Battle Creek, Mich.: Veterans Administration, 1966.

TAYLOR, J. A., & MAHER, B. A. Escape and displacement experiences as variables in the recovery from approach-avoidance conflict. *Journal of Comparative and Physiological Psychology,* 1959, **52**, 586–590.

TAYLOR, J. A., & RENNIE, B. Recovery from approach-avoidance conflict as a function of escape and displacement responses. *Journal of Comparative and Physiological Psychology,* 1961, **54**, 275–278.

WEISS, S. J., & MONAHAN, J. Matched-subject comparison of the toleration and exhaustion methods of extinguishing avoidance behavior. *Proceedings of the 77th Annual Convention of the American Psychological Association,* 1969, **4**, 87–88. (Summary)

WILSON, E. H., & DINSMOOR, J. A. Effect of feeding on "fear" as measured by passive avoidance in rats. *Journal of Comparative and Physiological Psychology,* 1970, **70**, 431–436.

WOLPE, J. Objective psychotherapy of the neuroses. *South African Medical Journal,* 1952, **26**, 825–829.

WOLPE, J. *Psychotherapy by reciprocal inhibition.* Stanford: Stanford University Press, 1958.

WOLPE, J. Isolation of a conditioning procedure as the crucial therapeutic factor: A case study. *Journal of Nervous and Mental Disease,* 1962, **134**, 316–329.

WOLPE, J. *The practice of behavior therapy.* New York: Pergamon Press, 1969.

(Received January 16, 1970)

Psychological Bulletin
1971, Vol. 76, No. 1, 15–38

ETHNIC STEREOTYPES [1]

JOHN C. BRIGHAM [2]

Florida State University

Present conceptualizations of stereotypes in psychology are outlined, and theoretical approaches linking ethnic stereotypes to negative attitudes and to discriminatory behavior are discussed. Research results concerning the content of ethnic stereotypes, their development and resistance to change, and the reciprocal nature of intergroup stereotypes are reviewed. Several critical issues, such as methodological problems and the "kernel of truth" controversy, are also discussed. It is suggested that none of the current definitions of stereotype are completely adequate. It is proposed that much of the ambiguity, both conceptual and empirical, in this area may be reduced if stereotypes are regarded as generalizations that are considered as unjustified by the person who affixes the label. There are several types of criteria under which a generalization may be considered unjustified. Possible research implications of this approach are presented.

One of the terms most frequently employed in research and theorizing concerning ethnic attitudes, prejudice, and intergroup perception and conflict is *stereotype*. The importance of this concept and the intellectual processes presumed to underlie it have been emphasized by a plethora of research concerned with ethnic attitudes and intergroup relations.

Yet, almost 50 years after the term stereotype was coined, there is little agreement between researchers and theorists concerning the function and importance of ethnic stereotypes in social perception and behavior. At a more basic level, there is little consensus on just what a stereotype *is*. Although much, perhaps most, "stereotype research" has proceeded with scant attention paid to such ambiguities, it seems to this writer that the accumulated research products will necessarily remain fragmented and of limited value until these conceptual questions are resolved, or at least recognized.

The conceptual work of Lippmann (1922) and the methodology developed by Katz and Braly (1933) have served as models for most of the subsequent theorizing and research in this area. Lippmann's original definition of stereotypes as being factually incorrect, produced through illogical reasoning, and rigid was sufficiently broad that most subsequent researchers have simply identified *one* of the above characteristics as the hallmark of an ethnic stereotype. This emphasis on differing criteria of stereotypes described by Lippmann, and the implications that accompany each of them, have led to considerable variance in the current conceptualizations of stereotype, as is pointed out in upcoming portions of this paper.

Given this situation, one might expect that there would have been several distinct paths of stereotype research in psychology, each attempting to gather data on one of the characteristics of stereotypes described by Lippmann. However, most researchers, regardless of their theoretical orientation, have utilized the Katz and Braly (1933) research paradigm. It is suggested later in this paper that the data obtained from this paradigm have little direct bearing on *any* of the conceptual criteria presented above. It would seem appropriate, therefore, to take a fresh look at the concept of ethnic stereotype and the relevant data that have been gathered in the past 40 years.

The results of empirical investigations of stereotypes and stereotypers are therefore reviewed. Since the most commonly investigated object of stereotypes has been Negroes, stereotypes of Negroes are used as a continuing point of reference in this review. In addition, particular attention is paid to several issues that are of considerable current interest, such as whether ethnic stereotypes must contain at least a "kernel of truth" in order to endure over time, and theoretical and empirical work dealing with the relationship of ethnic stereotypes to prejudicial attitudes and discriminatory behavior. After such an overview of the area has been obtained, it would seem appropriate to attempt to "pick up the pieces" and clarify the meanings, implicit and explicit, of ethnic stereotypes. In brief, such a clarification points to the necessity for recognizing that the values and standards of the researcher or theoretician are inseparably involved in his classification of certain beliefs or generalizations as stereotypes. While such a recognition does *not* mean that the concept is stripped of its utility, it does imply that the standards of the researcher be made more explicit. Such an enumeration of these standards, it is suggested, can lead to potentially productive paths of meaningful research. In line with this suggestion, several research directions which seem to hold promise are described.

---

[1] This paper was supported, in part, by Training Grant MH 06613-11 in Social-Personality from the National Institute of Mental Health to the Psychology Department, University of Colorado. The author would like to express his appreciation to Keith E. Davis, Stuart W. Cook, William W. Haythorn, and Jack E. Hokanson for their suggestions and comments on earlier drafts of this paper.

[2] Requests for reprints should be sent to John C. Brigham, Department of Psychology, Florida State University, Tallahassee, Florida 32306.

DEVELOPMENT OF THEORY AND METHODOLOGY

Origin of the Term

Credit for introducing and delineating the concept of stereotype goes to Walter Lippmann. In his 1922 book *Public Opinion*, Lippmann was concerned with the "quasi-environment" which is inserted between man and his real environment. In order to deal with a world which is too complex for direct acquisition, man constructs a "picture inside his head" of the world beyond his reach. The form of this "picture" is at least partially culturally determined; that is, "we pick out what our culture has already defined for us, and we tend to perceive that which we have picked out in the form stereotyped for us by our culture [Lippmann, 1922, p. 81]."

Lippmann posited that stereotypes are undesirable with regard to their largely incorrect contents and also as products of a faulty reasoning process. The process represents a type of short-circuiting that "imposes a certain character on the data of our senses before the data reach the intelligence [p. 98]." Lippmann reserved his strongest criticism fo the content of the stereotype, however. H spoke of the "strange connection" betwee the stereotype and the facts, and the necessit that the stereotype and the facts must "always" part company. Finally, Lippmann emphasized the rigidity of stereotypes and thei resistance to education.

The Katz and Braly Paradigm

The most frequently cited study of ethni stereotypes remains that carried out a Princeton in 1932 by Katz and Braly (1933) Considerable attention is paid to this investigation since, along with being the first empirical study of ethnic stereotypes,[3] the paradigm created by these researchers has serve as a direct model for the majority of subsequent stereotype research.

One hundred Princeton students were aske to list the traits that they considered mos characteristic of 10 ethnic groups: Negroes Germans, Jews, Italians, English, Irish Americans, Japanese, Chinese, and Turks Subjects chose the traits from a list of 8 traits which they were given. They were aske to

> Read through the list of words . . . and select thos which seem to you to be typical of the particula ethnic group. Write as many of these in the followin spaces as you think are necessary to characteriz these people adequately. If you do not find prope words . . . for all the typical ethnic group characteristics, you may add those which you think necessar for an adequate description [Katz & Braly, 1933, p 282].

After this procedure had been followed for al 10 ethnic groups, subjects were asked to g back and star the five words on each lis that seemed the most typical of each grou Only the starred items were used in the dat analysis.

---

[3] The work of Rice (1926) and Litterer (1933) while antedating the Katz and Braly study, was concerned with occupational stereotypes, rather tha stereotypes of ethnic groups. The term "ethnic" i used herein to denote groups that are of differin national origin (e.g., Turks, Germans), a group tie together by a common religion and cultural heritag (Jews), and a group defined by racial characteristic (Negroes).

The subjects showed a very high degree of agreement on the assignment of traits to ethnic groups. For example, Negroes were characterized as "superstitious" by 84% of the subjects and as "lazy" by 75%; Jews were seen as "shrewd" by 79%; and Germans were viewed as "scientifically-minded" by 78%. A measure of "definiteness" of the stereotypes was arrived at by calculating the least number of traits that would have to be included to account for at least 50% of the 500 "votes" cast by the 100 subjects for each ethnic group. Given the list of 84 traits from which to choose, the theoretical range is from 2.5 (perfect agreement by everyone) to 42.0 (perfect disagreement); the obtained range was from 4.6 traits (Negroes) to 15.9 traits (Turks). Thus the subjects' categorizations showed considerable agreement with each other.

Katz and Braly concluded that definiteness of stereotypes bore no relationship to liking for an ethnic group or prejudice toward it, since the most disliked groups—Negroes and Turks—were the most and least definitely stereotyped groups. This tendency toward a consensus in the categorization of ethnic groups was seen by the authors as evidence of fallacious thinking: "Stereotyped pictures of racial and national groups can arise only so long as individuals accept consciously or unconsciously the group fallacy attitude toward place of birth and skin color [Katz & Braly, 1933, pp. 288–289]."

THEORETICAL APPROACHES TO STEREOTYPING

Meaning of "Stereotype" in Psychology

Incorrect generalizations. Generalizations about members of a category (in this case, an ethnic group) can be construed as incorrect in two general senses. The first possibility for error is in *directionality,* for instance, the generalization that most members of Group X possess Attribute Y when, in fact, most members of Group X do *not* possess Attribute Y. The second possibility for error is in the magnitude of the generalization, for instance, the generalization that "almost all" (say, 95%) of the members of Group X possess Attribute Y, while in reality only "most" members (say 65%) of Group X possess Attribute Y. It would seem that this latter type of error is what is usually referred to by the term "overgeneralization."

Katz and Braly (1935) proposed a view of stereotypes that has persisted, in some quarters, to the present day. They stated: "A stereotype is a fixed impression, which conforms very little to the facts it tends to represent, and results from our defining first and observing second [p. 181]." This position is shared by Klineberg (1951) and Prothro and Melikian (1954).

A larger group of authors is content to equate stereotype with generalizations, which, although overgeneralizations, may still describe situations that are extant, although to a lesser extent than that expressed in a stereotype. A number of authors propose such a view—among them Schoenfeld (1942), Zawadzki (1948), Bogardus (1950), Allport (1954), Wedge (1966), Krech, Crutchfield, and Ballachey (1962), Gilbert (1951), Campbell (1967), and Rokeach (1968).

There is one major difficulty with definitions such as those outlined above. Characterizing something as invalid necessarily implies that a validity criterion is available. Yet, with reference to the actual distribution of traits within ethnic groups, empirical evidence is exceedingly scanty. Thus, in most cases, no criteria are available for assessing the factual validity of an ethnic generalization. Certainly, if an ethnic generalization is made in *exceptionless* terms, knowledge of the range of individual differences alone would seem sufficient grounds for terming such a generalization invalid. If, on the other hand, the generalization is phrased in terms of "most" or "some," considering the present lack of relevant data, the validity of this generalization cannot be assessed.

Generalization of unspecified validity. Many researchers treat the issue of the validity of generalizations as irrelevant (Abate & Berrien, 1967; Blake & Dennis, 1943; Brown, 1958, 1965; Child & Doob, 1943; Diab, 1962, 1963a; Edwards, 1940; Sherif & Sherif, 1969; Sinha & Upadhyay, 1960a). In this view, any generalization is a stereotype.

Others (Fishman, 1956; Mann, 1967a, 1967b; Mead, 1956; Schuman, 1966; Vinacke, 1949, 1956) do not consider validity irrelevant, but recognize that classing stereotypes as *over*generalizations implies the existence of a criterion of validity. However, as already indicated, such a criterion has seldom been available.

Categorizations and concepts. Some writers speak of stereotypes as the use (or misuse) of categories. This orientation is expressed by Secord (1959): "A stereotype is commonly thought of as involving a categorical response; i.e., membership in a category is sufficient to evoke the judgment that the stimulus person possess *all* [italics added] the attributes belonging to that category [p. 309]." Similar orientations are expressed by Harding, Kutner, Proshansky, and Chein (1954), Sanford (1956), Saenger and Flowerman (1954), and Simpson and Yinger (1965).

When stereotypes are treated in this manner, the category becomes isomorphic with the observer's *concept* of the group. Vinacke (1957) has proposed that "stereotypes should properly be regarded as concept-systems, with positive as well as negative functions, having the same general kinds of properties as other concepts, and serving to organize experience as do other concepts [p. 229]." Concepts are seen as "cognitive organizing systems" which, in conjunction with attitudes, operate in the control of response.

Concepts have both extensional properties —those properties corresponding to the objective characteristics of the object, that is, those upon which nearly every normal person would agree; and intensional properties—those features that result from a person's unique experience with objects, which make the concept personalized and which make it differ from the concepts of other persons. Stereotyping is a conceptual process, Vinacke proposed, whose crucial aspect is that personality traits are involved in the concept, as well as physical characteristics, and that the former operate as if they were *extensional* properties. Seen in this way, stereotypes can be accepted as an inevitable consequence of social learning, as the means whereby kinds of objects (persons) are classified on the basis of perceived properties, thus facilitating meaningful response to those objects. Presumably, such properties are incorporated into the concept in a probabilistic manner, so that exceptions, that is, persons whose attributes do not match all of those included in the concept, can be recognized.

Products of a "faulty" thought process. If a stereotype is looked upon as a generalization that may or may not be valid, then the question arises as to what distinguishes a stereotype from an "ordinary" generalization. To many psychologists, the crucial difference is the way in which the stereotype is arrived at. This viewpoint expands upon Lippmann's (1922) point that the stereotyping process is a type of short-circuiting, which "precedes the use of reason [p. 98]."

There appear to be two variants of this "faulty process" view. In the first sense, the process *itself* is seen as inferior, for instance, the person is using nonlogical reasoning. The alternate view is that the person's *basis of acquisition* falls below acceptable standards, that is, he develops his beliefs from sources that should not be employed, such as hearsay. Klineberg (1951) stated this latter position as follows: "Unlike other generalizations . . . stereotypes are based not on an inductive collection of data, but on hearsay, rumor, and anecdotes—in short, on evidence which is insufficient to justify the generalization [p. 505]."

The view that stereotypes are the result of an inferior process itself has been supported by Sanford (1956), Zawadzki (1948), and Bogardus (1950). Fishman (1956) presented this outlook in straightforward form:

> one man's objectively and analytically determined facts can be another man's stereotypes; in fact, one man's stereotypes may, subsequently, become the same man's analytically derived facts. Thus it is not the data which necessarily undergoes a change, but the process of arriving at the data [p. 34].

It should be mentioned that writers who have taken the position that stereotypes are invalid overgeneralizations also may be in congruence with the "inferior process" position. Presumably, if a person has *himself* developed an invalid overgeneralization, this will have been accomplished either through some sort of faulty thought process or from an unacceptable basis of acquisition. (It is, of course, also possible that his generalization has been learned from a respected source.) It does *not* necessarily follow, however, that the results of a faulty thought process *must* themselves be invalid (Fishman, 1956; Triandis & Vassiliou, 1967). It should be mentioned that although data on the process of acquisition should be obtainable, little work has been directed toward this end. Thus, as is the case with definitions stressing factual incorrectness, theorists are generally unable to present data that support their contentions.

Generalizations characterized by rigidity. Still another criterion that has been used for the identification of stereotypes is their rigidity. To Kerr (1943), the hallmark of stereotypes is that they "do not develop but remain ossified [p. 37]." Meenes (1943), Bogardus (1950), Adorno, Frenkel-Brunswik, Levinson, and Sanford (1950), Sanford (1956), Simpson and Yinger (1965), Katz and Braly (1935), Katz and Stotland (1959), Sinha and Upadhyay (1960a), Rokeach (1960), and Scott (1965) have also stressed rigidity.

Fishman (1956) brought up a point that is relevant to those descriptions of stereotypes that stress their rigidity. He pointed out that a stereotype can be seen as rigid only if it is not modified in the face of (*a*) new information; (*b*) changes in a person's needs, motives, or interests; or (*c*) a changed focus of interaction. If there is not change in any of these areas, then there is no rigidity implied when a stereotype remains unchanged. Thus, "rigidity *may* be an aspect of some stereotypes (and of some attitudes in general) but it *must* not be of all [Fishman, 1956, p. 39]."

Stereotypes as habits. Campbell (1967) has attempted to analyze stereotyping from within the framework of Hullian theory. Essentially, Campbell is modifying a formula used to explain behavior into one to explain beliefs. Stereotypes, under this rubric, are regarded as "perceptual ${}_sE_R$'s [Campbell, 1967, p. 821]." Stimulus intensity, V, can be regarded as the *contrast* between groups. Since a greater value of V will lead to a greater ${}_sE_R$, it follows that the greater the real differences between groups on any particular custom, detail of physical appearance, or item of material culture, the more likely it is that the feature will appear in the stereotyped imagery each group has of the other.

The "intrapsychic" D, K, and H determinants are regarded as "projections" of the personality of the perceiver. The more remote and less well-known the outgroup, the more purely projective the content of the stereotype will be. Also, among trait differences of approximately equal contrast (equal V), those which are drive (D) and incentive (K) relevant and which are familiar (H) are most apt to be noticed and represented in stereotyped imagery. The implications of this approach are that most stereotypes will be somewhat valid (due to V). It is those cases in which the outgroup is relatively unknown where the stereotype may be almost completely invalid.

Why Are Stereotypes Undesirable?

The prevailing orientation of social scientists toward the use of stereotypes has been summed up by Brown (1958): "They think it is at least irrational and probably wicked to subscribe to them . . . [p. 364]." As reviewed above, most theorists have regarded stereotypes as "wrong," either because they do not correspond to the facts about the ethnic group, or because they were arrived at through some unacceptable process. Yet, as already mentioned, criteria for assessing the degree to which stereotypes have met these standards (content and/or process) are largely unavailable. Thus, such formulations must remain at the level of assumptions. The widespread adoption of such assumptions has led Brown (1958) to ask, "Is it possible that the social psychologist has used the word *stereotype* to stigmatize beliefs of which he disapproves but which he does not know to be false? Has he perverted his science to achieve a moral purpose [p. 366]?"

Brown and Campbell have attempted to go beyond the views discussed thus far, to isolate those characteristics of stereotypes that are undesirable. Brown (1958) proposed that the "intellectual respectability" of the popular stereotypes has been "convincingly" impugned. Stereotypes, he stated, are often not well-founded in direct experience; they sometimes serve to rationalize selfish behavior; they are often not sensitive to contrary evidence; and they ascribe to racial inheritance that which may be a cultural acquisition. In a later work, Brown (1965) was content to summarize the objectionable aspect of stereotypes as their ethnocentrism and emphasis on cultural absolutism, the implication they convey that ethnic groups have inborn and unalterable psychological characteristics.

Campbell (1967) listed four general characteristics that he felt were wrong with stereotypes. The first point is the "phenomenological absolutism" of the ingroup member's imagery of the outgrouper or minority group member. Naïvely, one assumes without question that the outgroup is as one perceives it, or as the ingroup informs one about it. The second deficiency is in the amount of between-group difference perceived, the exaggerated homogeneity attributed to groups, and the nonrecognition of the amount of overlap between any two groups.

The third point concerns erroneous causal perception. The stereotyper is liable to attribute group differences to racial, rather than environmental causes. This point is very similar to Brown's objection discussed above. Campbell's fourth point also is directed toward a faulty conception of causality, this time concerning the hostility felt toward the outgroup. The stereotyper believes that it is undesirable characteristics of the outgroup that cause his hostility, rather than recognizing that it is his preexisting hostility that has caused all possible between-group differences to be interpreted in terms of the despicable characteristics inherent in the outgroup. It is this last factor, Campbell believed, which is "the most important although the most difficult to explicate [Campbell, 1967, p. 825]."

Brown and Campbell, then, focus directly on the characteristics of the stereotyp*er*—his unrecognized hostility, ethnocentrism, and naïveté. With regard to the generalization itself, attention is given to its emphasis on racial rather than environmental causes, and the exaggerated amount of between-group differences perceived.

RESEARCH FINDINGS ON ETHNIC STEREOTYPES

Incidence and Content

A voluminous amount of research effort has been directed toward the assessment of stereotypes held by members of one national group about the members of selected other national groups. Most researchers have used the Katz and Braly experimental framework *in toto,* often using the identical 84 adjectives (translated when necessary). Simple recitation of the results of such research can serve little purpose here. The reader is directed to the articles themselves (see, e.g., Bastide & van Den Berghe, 1957; Chandra, 1967; Das, 1962; De Bie, 1951; Jahoda, 1959; van Den Berghe, 1962; Zaidi & Ahmed, 1958) for specific results.

Stereotypes of Negroes as assessed by the Katz and Braly paradigm have tended to be rather definite and quite negative. Such traits as superstitious, lazy, dirty, and ignorant usually have headed the list in frequency of attribution. Recent psychological research and public opinion polling results indicate that although the United States white's view of Negroes may be becoming somewhat less negative, it is still radically different from his view of white Americans. It should be added that this view of Negroes is not limited to the United States. Subjects in India, Lebanon, Great Britain, Pakistan, and other countries hold similarly negative stereotypes of Negroes (e.g., Prothro & Melikian, 1954). There is evidence that many Negroes accepted this view also.

Bayton (1941) tested Negro college students at Virginia State College. Using the Katz and Braly technique, subjects were told to check the most typical traits for 10 groups, including "the Negro." Among the traits most commonly assigned to "the Negro" were musical, superstitious, very religious, lazy, loud, and happy-go-lucky. When asked to describe the "typical Virginia State College student," however, the subjects checked talkative, intelligent, lucky, sportsmanlike, jovial, sophisticated, and other more favorable traits.

Apparently, the description "the Negro" implied "the lower-class Negro" or "the average Negro, who is of lower-class" to these subjects (and perhaps to many white subjects in other studies also). Bayton commented:

> Here is a group accepting what has been called the "group fallacy." Also, they accept, without contradiction, the stereotypes assigned other minorities as

> shown in their characterization of Jews. Moreover, they largely accept the stereotype usually assigned to themselves . . . [p. 102].

Jewish high school students also have tended to accept the prevailing stereotype of Jews (Dewey & Humber, 1966, p. 363).

The importance of offering additional items of descriptive information is clearly demonstrated in a study carried out by Bayton, McAlister, and Hamer (1956). These investigators differentiated the object groups by race (Negro-white), but also by social class (upper-lower). It was found that the resulting characterizations varied more as a function of class than of race.[4] Thus upper-class whites *and* Negroes were seen as intelligent, ambitious, industrious, neat, and progressive, while lower-class whites and Negroes were assigned the traits, ignorant, lazy, loud, and physically dirty, in common. Lest these findings be taken as evidence of the absence of racial stereotypes, however, the authors caution:

> It is altogether possible that when the subjects are asked to stereotype "white Americans" and "the Negro" they assume that *most* white Americans tend toward the upper-class positions while *most* Negroes tend toward the lower-class status [p. 84].

Development of Ethnic Stereotypes and Their Resistance to Change

Although there has been little research directed toward analyses of the processes through which stereotypes arise, that which has been carried out has focused on Negroes as the object group. Skin color and racial membership appear to be perceived as possessing both racial and social significance by children by the time they reach age 4 or 5 (Clark & Clark, 1939; Goodman, 1952; Morland, 1962). Numerous studies have indicated that white skin color is overvalued and dark skin color undervalued by both white and Negro children (e.g., see Asher & Allen, 1969; Goodman, 1952; Greenwald & Oppenheim, 1968). But few researchers have investigated whether this early racial awareness and valuation is accompanied by the differential attribution of stereotype-relevant traits to whites and Negroes.

Taylor (1966) showed nine pictures of a Negro and a white child engaged in some activity to children aged 6 to 10. For each picture, they were asked to identify which person had a given negative stereotype-relevant trait. He found that white children assigned the negative characteristics significantly more often to the Negro child in the photograph than to the white. In addition, Negro children assigned two of the traits (dirty and dumb) significantly more often to the Negro child than to the white child.

An interesting analysis of the development of stereotypes in school children was carried out by Blake and Dennis (1943). White schoolchildren in Charlottesville, Virginia, were to check whether each of 60 traits was more common in the Negro, in the white, or "no difference." Subjects in Grades 4 through 11 were tested.

The authors calculated a Preponderance Index, which was simply the percentage of subjects who assigned a trait to Negroes minus the percentage assigning it to whites. The traits most overwhelmingly attributed to Negroes were superstitious, move slowly, easily frightened, better servant, flashy clothes, ignorant, sleepy-natured, and happy-go-lucky (all Preponderance Indexes above 80). Interestingly, the Preponderance Indexes for these and other traits increased dramatically for the older children (Grades 8 through 11). It appeared that the young white child first acquired a generally unfavorable attitude toward the Negro. It was only as he grew older that he gradually learned to apply the "adult stereotype." Thus, the older whites applied traits to the Negro that were a part of the stereotype, but which were not in and of themselves uncomplimentary, such as cheerful, fond of music, religious, sense of humor, sense of rhythm, and less selfish. The younger children made no such differentiations and applied negative traits more indiscriminately to the Negro.

Unfortunately, longitudinal studies of the degree of stability or change of individual stereotypes over time have not been carried out. Several studies have indicated, however, that the modal group stereotypes can show

---

[4] Bruner and Perlmutter (1957) have also demonstrated the importance of additional items of information in studies of "stereotyping."

Psyl. Research
6·3·81

considerable change in the face of changing external situations. Thus, Dudycha (1942) and Meenes (1943) found that traits assigned to Germans and Japanese by Americans were more negative during World War II than they had been in pre-World-War-II studies. Sinha and Upadhyay (1960b) found that students in India assigned significantly more negative traits to Chinese a few months after the 1959 Chinese-Indian border dispute than they had a few months before the dispute.

A series of studies has indicated that the traits most commonly attributed to ethnic groups show considerable stability over time. Centers (1951), for example, took the 12 most frequently assigned traits for each ethnic group as found by Katz and Braly in 1932, and in 1950 asked subjects to match the adjective lists with the ethnic group they represented. For all 10 ethnic groups, over 70% of the subjects made the correct match. Hoult (1953) was unable to replicate these results, however.

Using the Katz and Braly technique, Gilbert (1951) compared group stereotypes obtained from Princeton students in 1950 with those obtained by Katz and Braly in 1932. Although the ordering of characteristics varied somewhat, the traits most commonly assigned to the various groups in 1932 were, for the most part, assigned again in 1950. The students in 1950, however, showed less agreement among themselves about assigned traits, that is, the overall group stereotypes were less "definite." Gilbert interpreted this finding as "impressive evidence of *resistance* to the stereotyping tendency, and a *fading-out* of such formulations [Gilbert, 1951, p. 246]."

A third investigation of the trait attributions of Princeton students has recently been carried out by Karlins, Coffman, and Walters (1969). (These studies might now be termed the Princeton trilogy.) Using the identical procedures, they found that while the frequency of some "traditional" trait assignments had declined, other "new" traits had arisen, for which there was considerable intersubject agreement in attributions to specific groups. In fact, although the degree of intersubject agreement on trait attributions was less than had been found by Katz and Braly (1933), it was considerably greater than that found by Gilbert (1951). Nevertheless, Karlins et al. posited that their findings indicate that "younger generations show more careful thinking about ethnic generalization than their counterparts of the 1930's [p. 14]."

Self-Stereotypes and Reciprocity of Ethnic Stereotypes

While a group of subjects (themselves an ethnic group) may have stereotypes about another ethnic group, it follows that members of the second ethnic group will also have stereotypes about the first. This two-way reciprocal feature of ethnic stereotyping has been investigated by Vinacke (1949, 1956), Buchanan (1951), Riffenburgh (1966), Schuman (1966), Triandis (1967), Campbell (1967), and Abate and Berrien (1967). Researchers working with multiple ethnic groups have found a significant amount of agreement between different ethnic groups' views of any specific rated group. This consistency has been found in Hawaii (Vinacke, 1949), Europe (Buchanan, 1951), and Pakistan (Schuman, 1966).

Another possible product of cross-cultural research is the comparison of Group A's stereotype of Group B with Group B's characterization of themselves (self-stereotype). Campbell (1967) and LeVine (1965) have examined stereotype data from 50 respondents in each of 30 tribal or linguistic regions in East Africa, each region giving "attitudes" toward 13 outgroups. They found that in many cases there was considerable agreement between neighboring groups and the target group. As might be expected, it has been found that traits regarded as typical of one's own group are usually "overvalued," that is, they are thought to be more favorable or complimentary by the groups than other ethnic groups who do not consider the trait typical of themselves (Vinacke, 1956).

Abate and Berrien (1967) have carried this paradigm one step further by examining stereotypes, self-stereotypes, and the estimated "real" characteristics of the group (which they term the "vereotype"), as measured by the Edwards Personal Preference Scale (EPPS). Their subjects, Japanese and American students, were presented with 15 behavioral orientations drawn from the EPPS and asked to rate, on a 5-point scale, how well each description "fitted" Japanese or Americans in general. Mean self-report EPPS scores constituted the "vereotype."

Rank-order correlations demonstrated that each group was quite accurate in reproducing the other group's "self-stereotype"; but the correspondence between stereotypes and "vereotypes" was less impressive. The correlation between Japanese stereotypes of Americans and the American vereotype was .54, but the correlation between the American stereotype of Japanese and the Japanese vereotype was .20 for women and −.04 for men. Surprisingly, the respondents were no more accurate in perceiving the vereotype of their own national group than in perceiving the vereotype of the foreign group. American self-stereotypes agreed with their vereotypes to a much greater extent than was true for the Japanese subjects, however.

The question arises as to whether the EPPS serves as a suitable criterion for the "true" characteristics (i.e., vereotype) of the groups in question. This situation is the obverse of one in which judgments of some sort (e.g., peer ratings) are used as the validation criteria and the instrument (e.g., the EPPS) is compared with these criteria. Certainly, *if* the EPPS "vereotypes" are seen as an acceptable validation criterion, then it would seem that the proposition that national stereotypes represent the recognition of true group differences (Campbell, 1967) does not hold in all cases. Of course, it should be noted that Campbell was speaking of interacting groups—tribal or linguistic regions in East Africa. Presumably, the two national groups used by Abate and Berrien had had little or no opportunity to interact with one another.

Relationships among Attributes of Stereotypes

Edwards (1940) has proposed four "dimensions" of stereotypes that might profitably be analyzed: uniformity, direction (favorable or unfavorable), degree of favorableness, and quality (actual content). Uniformity refers to the extent to which the individual's response is in accord with the responses of others (viz., earlier discussion of the "definiteness" of stereotypes). Under this rubric, then, most studies of stereotypes have dealt with the uniformity dimension.

The evidence seems to indicate little or no relationship between uniformity of ethnic stereotypes and stated preference for the ethnic group. Taft (1959) has calculated rank-order correlations for some of the relevant published data, with the following results: The uniformity-preference correlation in the Katz and Braly (1933, 1935) studies was .07, in Schoenfeld's (1942) study it was .19, while Gilbert's (1951) data produced a −.01 correlation. Although their data are not quantifiable, Vinacke (1956) and Prothro and Melikian (1954, 1955) found a similar lack of relationship. Taft's own data, with familiarity partialed out, yield preference-uniformity correlations of .11 and .01.

Although rated preference for the group seems to show little relationship to uniformity, favorability of the actual traits assigned does appear to be related to stereotype uniformity. Using ratings of the "degree of favorableness" of the trait adjectives, gathered after subjects had completed the trait-attribution task, Karlins et al. (1969) computed overall favorableness ratings for each ethnic group. Each trait's frequency was multiplied by its mean favorableness value, and the sum of the values was divided by the total frequency. They found that favorableness showed a .77 correlation with uniformity in their sample. Using the same favorableness ratings, Karlins et al. calculated favorableness-uniformity correlations of .16 and .68 for the data of Katz and Braly (1933) and Gilbert (1951). Further, rank orderings of ethnic groups according to favorability scores derived from the above two studies correlated .75 and .52 with those derived by Karlins et al. (This finding is valid *only* to the extent that the traits have remained constant in their perceived favorableness over this period, of course.) These findings, contrasted with those of Taft (1959) discussed above, suggest that mean *favorableness* of stereotype about an ethnic group is *not* a good indicator of *preference* for that group, at least when these are averaged over all subjects.

One attempt has been made to relate the extent to which *individual* subjects agree with the modal patterns of trait attributions, to attitudes toward the ethnic group. Working on the assumption that responses from 200 college students will yield a rough representation of the (college) subsample's modal stereotypes assessed under the Katz and Braly paradigm, Brigham (1969) calculated a score denoting the degree to which each subject's five traits assigned to Negroes agreed with the most common assignments made by the group as a whole. This score was then correlated with score on a much-validated measure of attitudes toward Negroes. No relationship was found ($r = .03$), although subjects' agreement with modal trait assignment to (presumably white) Americans was slightly correlated ($r = .14$, $p < .05$) with anti-Negro attitude.

Familiarity with the ethnic group does appear to show a nonambiguous relationship to uniformity of stereotype, however. In Taft's reanalysis, Katz and Braly (1933), Schoenfeld (1942), and Gilbert (1951) obtained familiarity-uniformity correlations of .51, .20, and .66, respectively. (It should be noted here that in Schoenfeld's own analysis of his data, he did not feel that the familiarity-uniformity relationship was significant.) For Taft's data, the familiarity-uniformity correlation was .85. Triandis and Vassiliou (1967), on the other hand, found that while American stereotypes of Greeks became more uniform with contact, the relationship did not hold for the Greek stereotype of Americans.

The familiarity-preference interaction was also investigated by Taft. While he found a very high positive relationship ($r = .92$), previous studies did not yield such striking results. As Scott (1965), Triandis and Vassiliou (1967), and others have pointed out, the nature of the relationship would seem to depend upon the circumstances under which familiarity is acquired (e.g., friendly cooperation or war), and the kind of interdependence that the groups maintain.

Vinacke (1957) has suggested that findings on trait attributions to the ingroup (self-stereotypes) be considered as another facet of stereotyping. As might be expected, the traits assigned to one's own group tend to be more favorable than those attributed to outgroups (Campbell, 1967; Child & Doob, 1943; Karlins et al., 1969; Schuman, 1966; Sherif & Sherif, 1969; Vinacke, 1949, 1956). Furthermore, as mentioned earlier, traits regarded as typical of one's own group are usually "overvalued."

To summarize the findings in this area, it appears that when data are summed over all subjects, the uniformity of stereotypes is related to favorableness of the most commonly included traits, but *not* to preference for the ethnic group. Individuals' agreement with the most common patterns of trait attributions does not appear to be related to attitude toward the ethnic group, at least when Negroes are the object group in question. Uniformity of stereotype seems to be positively related to familiarity with the object group, and traits attributed to one's own group tend to be more favorable than traits applied to other groups. The relationship between familiarity (or degree of contact) and preference would seem to be dependent upon the circumstances under which the familiarity was achieved.

CRITICAL ISSUES IN STEREOTYPE RESEARCH

Methodology

There are several recurring methodological weaknesses that have characterized much of the stereotype research reviewed thus far. Probably the most obvious problems are inherent in the Katz and Braly technique, which has been so widely utilized. This method, it will be remembered, directs the subject to pick out (from a list he is given) those adjectives that he considers to be the "most typical" of each ethnic group. This procedure, then, may *force* the subject into thinking in terms of generalizations and categories.

Eysenck and Crown (1948), in a study of national stereotypes employing the Katz and Braly technique, found that 19% of their subjects refused to take part in the project, declaring that it was meaningless or impossible. Of the remaining subjects, over half of them recognized that they were being forced to use stereotyped answers they had heard elsewhere. Eysenck and Crown concluded:

> our results suggest per contra that the majority of respondents have no particular views on the subject at all . . . when we look for stereotyped views, and give the subject no chance to reply in any non-

> stereotyped fashion, we should not be surprised that the answers we get are stereotyped ones. It is strictly inadmissible to conclude from such results, however, that the views actually held by the subjects are stereotypes . . . [p. 35].

Brown (1965) reminds us that the

> investigators of ethnic stereotypes did not give their subjects the opportunity to be sensible men. They forced decisions on the basis of one item of information—ethnic membership. May not their subjects have been making the best of such a task [p. 179]?

Ehrlich and Rinehart (1965) compared responses given to a Katz and Braly type of checklist to those given in an open-ended format. In the open-ended format, subjects listed about 5 traits per ethnic group, while they checked from 8 to 19 traits per group under the checklist condition. Differences between mean number of traits listed for every ethnic group under the two conditions were significant. Furthermore, although subjects in the checklist condition were told that they could add traits to the list, only 14 traits were actually added (0.2% of the total trait assignments made).

Even the number and kind of ethnic groups included on a list seem to have a significant effect on the picture that emerges of any specific ethnic group. Diab (1963a) found that modal trait attributions made about an unpopular ethnic group (Russians) became more unfavorable when that group's name was placed next to the name of a favored ethnic group (Americans). Further research (Diab, 1963b) has indicated that when the ethnic groups seen as most favorable by the subjects are omitted from the study, overall trait attributions made to groups that had previously been viewed with ambiguity become more favorable.

The Kernel of Truth Controversy

Can a stereotype continue to exist in a culture without any "kernel of truth" whatsoever inherent in it? Debate on this issue continues among those researchers concerned with ethnic stereotypes. The study most often cited by those who propose that stereotypes can exist *without* any kernel of truth was that of LaPiere (1936). In the 1920s in southern California, Armenian laborers were seen as (stereotyped as) dishonest, deceitful, liars, and troublemakers. In fact, LaPiere found that Armenians appeared less often in legal cases, applied less often for charity, and had as good credit ratings as those received by the other ethnic groups. While it would thus appear that the stereotype maintained itself without any supportive evidence at all, the necessarily superficial nature of this study seems to leave the issue somewhat in doubt.

Schoenfeld (1942) attempted to address the issue more directly with a study of "first name stereotypes." After a series of pretests, Schoenfeld selected eight masculine and eight feminine names for which there seemed to be wide agreement on the personality characteristics "belonging" to each name. In the final phase, subjects were told to match each name with the appropriate personality characteristics. For both male and female subjects, there was considerable agreement, well above chance levels. Interestingly, only 7 of 248 original subjects complained that the task was difficult or meaningless. Klineberg (1954) felt that this study "has demonstrated that stereotypes can and do develop without any basis in objective reality [p. 486]."

Wells, Goi, and Seader (1958) studied college students' characterizations of persons who buy different makes of new cars. In 1956, people who bought Plymouths were seen as "quiet, careful, slow, silent, moral, fat, gentle, calm, sad, thinking, patient, honest, content, and understanding." In 1957, however, Plymouth initiated a powerful advertising campaign, and students in 1957 (about 70% of whom had also been included in the 1956 sample) then described owners as "high class, important, rich, particular, different, and feminine [Wells et al., 1958, p. 121]." Thus, the image of Plymouth owners underwent a dramatic change in one short year.

That facts can affect stereotypes has been shown by those studies that focused on change due to external events, for instance, war. The similarities between stereotypes of one ethnic group held by various other ethnic groups have led Vinacke (1956) to postulate:

> It would be ridiculous to assert that groups of different national or cultural origin . . . do not have certain general characteristics which differentiate them from groups of different origin. . . . In part wrong, superficial, and limited, they [national stereotypes]

> nevertheless generalize some actual cultural traits [p. 285].

In a similar vein, Margaret Mead (1956) proposed that "the elements which make up such stereotypes are accurate but incomplete descriptions of national character in particular *cross-cultural* contexts [p. 222]."

The findings of agreement between stereotypes of a group and that group's self-stereotype would seem to lend support to the position that ethnic stereotypes *can* have validity (Campbell, 1967; Fischer & Trier, 1962; LeVine, 1965; Schuman, 1966; Vinacke, 1949, 1956). One possible precondition to such validity is suggested by Triandis and Vassiliou (1967) who, interpreting their finding with regard to Greek and American subjects, observed: "The present data suggest that there is a 'kernel of truth' in most stereotypes *when they are elicited from people who have firsthand knowledge of the group being stereotyped* [p. 324]."

On the other hand, Abate and Berrien (1967) have indicated that although there seems to be considerable agreement between stereotypes and self-stereotypes, there may be very *little* correspondence between stereotypes and modal group characteristics ("vereotype"), at least as measured by a personality inventory (the EPPS). In this connection, the problem of what type of *truth* one is interested in becomes relevant. If Group A has an image of Group B, which agrees, by and large, with Group B's image of itself, then can we call Group A's image incorrect if Group B's self-image appears largely incorrect with reference to psychological test results or ratings by independent observers? That is, can we expect Group A to have a *better* insight into the "true" characteristics of Group B than Group B has of its own characteristics? If Group A did indeed have greater insight, studies using self-stereotypes (by Group B) as validity criteria would conclude that Group A's stereotypes were invalid. Certainly, more work along the lines of the Abate and Berrien study would be valuable, especially if more powerful indexes of "true," existing group profiles can be developed and utilized.

In sum, there is ample evidence that ethnic stereotypes can have "kernels" of truth, at least in a convergent validity sense, that is, agreement between several groups as to the traits that characterize a particular object group. However, the absence of acceptable validity criteria leaves the question of whether *all* such stereotypes have some validity unanswered.

Relationship of Ethnic Stereotypes to Ethnic Prejudice and Discriminatory Behavior

Ethnic stereotypes and prejudice. Probably the most basic characteristic linking stereotypes and ethnic prejudice is an awareness of discrete groups. Obviously, in order to feel negatively toward a group, one must be able to perceive the different individuals of the given ethnic group as having certain constant characteristics, as being similar to other individuals in the same group, and as being different from individuals not of that ethnic group. In other words, in order to be prejudiced against Group A (and, in the same sense, to hold stereotypes about Group A), one must have a concept (category) of the group.

Several studies (Allport & Kramer, 1946; Elliott & Wittenberg, 1955; Himmelfarb, 1960; Lindzey & Rogolsky, 1950; Scodel & Austrin, 1957) have seemed to demonstrate that prejudiced persons are more *accurate* in their categorization of object-group members than are nonprejudiced subjects. These studies appeared to show that anti-Jewish subjects could more accurately identify photographs of Jews than could nonprejudiced subjects. However, with one exception (Himmelfarb, 1960), all the studies have found that anti-Semitic subjects judge *more* faces as being Jewish than do nonprejudiced subjects. It appears, as Secord and Backman (1964) proposed, that the anti-Semite errs on the "safe" side, putting a picture in the Jewish category if there is any likelihood that the person might belong there. It now appears that the enhanced accuracy earlier attributed to prejudiced persons was an artifact rising from their tendency to judge a greater number of faces to be Jewish, independent of the actual stimulus attributes.[5]

The undifferentiated nature of concepts (categories) used by prejudiced persons has been investigated by Secord and his co-workers. Secord, Bevan, and Katz (1956) selected 10 pictures of Negroes ranging along a rough continuum of "Negroidness" as judged by physiognomic variables. It was hypothesized that if the "overcategorization" factor of stereotyping is indeed operative, then there should be no difference in the personality characteristics assigned by the prejudiced subjects to the pictures varying in Negroidness, as long as they are perceived as belonging to the category (concept) "Negro." This was the case, as the judges were very similar in their attribution of traits to all the Negro pictures. The authors concluded that in the eyes of the high school subjects (in Atlanta, Georgia), "a 'Nigger' is a 'Nigger' no matter how white he may be [p. 60]."

Secord (1959) used a similar procedure. Subjects were given a list of 20 traits from which to choose; 4 had previously been identified as unfavorable aspects of the Negro stereotype, 4 related to favorable aspects of this stereotype, and 6 traits referred to favorable and 6 to unfavorable traits irrelevant to the cultural stereotype of the Negro. Again pictures of Negroes, varying in physiognomic Negroidness, were used. Subjects received a picture-stereotyping score, a verbal-stereotyping score, and a questionnaire-prejudice score. The subjects stereotyped all of the pictures to an equal degree, regardless of the amount of Negroidness, while the more prejudiced subjects stereotyped all of the pictures to a greater extent. Verbal stereotyping was related to prejudice to a greater extent ($r = .60$) than was picture stereotyping ($r = .36$).

The most widespread view among writers in this area is that stereotypes perform a *function* for the prejudiced person, allowing him to rationalize the hostility that he feels toward the group. Simpson and Yinger (1965), Gordon Allport (1954), Saenger (1953), LaViolette and Silvert (1951), and Ehrlich (1962a, 1962b) adopt such an orientation.

Other researchers regard stereotypes as an actual component, the "cognitive" aspect, of prejudicial attitudes (Harding et al., 1954; Harding, Proshansky, Kutner, & Chein, 1969; Katz & Stotland, 1959; Secord & Backman, 1964). Finally, Vinacke (1949) attempted a sort of synthesis of both of these orientations, stressing the "overlap" between stereotypes and prejudice. He proposed that stereotypes may serve as verbal expression of prejudice, or the rationalization or projection of it, or they may express a prejudice partly, or the stereotype may not be an expression of prejudice at all. Similarly, a prejudice may be apparent in a stereotype, may help determine it, may make use of the convenient labels, or may express itself differently, or be present without stereotyping (Vinacke, 1949, p. 288).

Longitudinal research into the development of negative attitudes and stereotypes, focusing on early childhood, might serve to shed light onto the actual relationship between stereotypes and prejudice. At present, empirical evidence on this question is relatively meager. Correlational analyses have indicated that ethnic prejudice and stereotyping seem to covary (Koenig & King, 1962, 1964; Saenger & Flowerman, 1954; Secord, 1959). Further, the ethnic categories of prejudiced persons appear to be less well-differentiated (Secord et al., 1956). While it seems to have been demonstrated that prejudice does not exist in the absence of stereotyping, little empirical attention has been directed toward investigating whether stereotyping can exist in the absence of prejudice.[6]

The widespread consensus that stereotypes function as "rationalizers" for prejudiced behavior seems intuitively sensible, but this analysis has been proposed and endorsed *without* grounding in empirical testing. Although the operationalization of a test of "rationalization" might be difficult to set up, it would seem that research in this realm would be eminently worth while.

---

[5] Perloe (1964) found that there was a tendency for subjects with strong *favorable* attitudes toward the stimulus groups (prep school or high school graduates) to be somewhat more accurate in their categorizations from pictures.

[6] It is clear that "positive" stereotypes, if the term is used in this sense, can exist in the absence of negative prejudice toward that group. However, only negative stereotypes are being considered in this discussion. Certainly stereotypes (as measured by the Katz and Braly technique) can exist in the absence of prejudicial *behavior* since subjects often express stereotypes toward groups with which they have not interacted, for instance, Turks.

Under the Katz and Braly paradigm, subjects have been willing (at least in the earlier studies) to stereotype virtually *all* ethnic groups. Unless some type of pervasive outgroup prejudice is postulated, it would appear that this indicates the presence of stereotyping without prejudice. It follows that these stereotypes are not functioning as rationalizers. The question then becomes, How and why do stereotypes become "attached" to prejudice to serve as rationalizers? Unfortunately, the dynamics underlying this question have not been elucidated.

Stereotypes and behavior. It has long been assumed that stereotypes have relevance to behavior, on an individual level and, perhaps, on an international level also. For example, in the series of resolutions adopted by the General Conference of the United Nations Educational, Scientific, and Cultural Organization (UNESCO) in 1947 and 1948 in connection with the Project on Tensions Affecting International Understanding, the Director General was authorized to promote a worldwide series of investigations of "national stereotypes" (Klineberg, 1951, p. 505).

While it seems logical that hostility can be generated more easily between two nations that hold unfavorable stereotypes regarding each other (Klineberg, 1951; Wedge, 1966), several of the studies already reviewed (Dudycha, 1942; Meenes, 1943; Sinha & Upadhyay, 1960a, 1960b) have seemed to indicate that stereotypes are not determiners of national policy, but tend rather to *reflect* this policy. Thus Buchanan (1951), in connection with the UNESCO project, stated:

> This alteration over a period of time suggests that stereotypes are less likely to govern the likes and dislikes between nations than to adapt themselves to positive or negative relationships based on matters unrelated to images of the people concerned [p. 526].

The crucial variable would seem to be whether the decision makers in the nations also held these stereotypes and whether they let the stereotype influence their decisions in particular cases. Unfortunately, data on these matters are not readily obtainable.

There is a similar paucity of data pertaining to the relationship between the stereotypes of individuals and discriminatory behavior toward members of the ethnic group. On the one hand, researchers interested in stereotypes usually have not employed behavioral measures. On the other hand, researchers concerned with actual behavior toward ethnic group members, typically Negroes, have characteristically employed only broad "attitude" or "prejudice" measures, whose relationship to stereotyping is unclear.

One attempt to assess the relevance of stereotypes to behavior (or, more specifically, to evaluations of individuals) has been made by Mann (1967a). He attempted to find out whether, when persons had attributed certain prevalences of traits to ethnic groups as a whole, their evaluations of specific ethnic group members would be consistent with the magnitude of their trait attributions for the entire group.

Mann hypothesized that it should be relatively "easy" for individuals to stigmatize an entire group as an "abstract mass." However, this blanket attribution of qualities would be more difficult in concrete individual cases. The outcome could be a readiness to see the ethnic group as "bad" but the individual ethnic group member as "not bad," even if the asserted prevalence of badness in the group would make it improbable that an individual group member lacks badness.

Working with white subjects in the Union of South Africa, Mann found that when it came to attributing traits to specific individuals within the ethnic groups, no subject was inconsistent with his generalization when it had been exceptionless (100%).[7] For those subjects whose generalizations had not been exceptionless (20%, 40%, 60%, or 80% of the ethnic group supposedly possessed the given trait), however, about two-thirds of the subjects showed some inconsistencies. As would be expected, most of these inconsistencies were for situations in which 40% or 60% of the group was estimated to have a particular trait. Contrary to Mann's hypothesis, however, most of the inconsistencies were ones in which a random individual was assigned a trait (usually negative) that the subject had previously assigned to a *minority* of the group. Only about half as many subjects denied the presence of negative traits in random individuals from a group to which they had assigned a high prevalence (60% or 80%) of the trait.

---

[7] Mann reported that 17% of the generalizations made by his subjects were exceptionless, that is, 100%. It could be expected that the prevalence of exceptionless generalizations about ethnic groups in the United States would be considerably less, contrasting the racial climates in the two nations. In fact, research carried out by this writer indicated that for United States college students in a somewhat similar task, the percentage of generalizations that were exceptionless (100%) was 2.2%. Viewed in another way, 42.5% of the subjects circled at least one "100%" in the 120 percentage-circlings they made.

Brigham (1969) attempted a further investigation of the extent to which expressed generalizations are significant factors in the evaluation of, and behavioral intentions toward, ethnic group members (Negroes). He found that some of the percentages circled regarding the prevalence of traits in "Negroes" as a group were significantly related to racial attitude and to average evaluations made of specific Negroes. However, generalizations made about "Negroes" on specific traits were *not* generally related any more strongly to evaluations of individual Negroes on that *same* trait than to evaluations of the individuals on other traits of similar affective loading. But while racial attitude showed a small significant relationship to behavioral intentions toward individual Negroes, an overall stereotyping measure showed no relationship whatsoever ($r = .04$) to behavioral intentions. This stereotyping score, however, *was* significantly related to the evaluation of individual Negroes, which, in turn, was significantly related to behavioral intentions toward the individual Negroes.

On the basis of the above two studies, at least, it would seem evident that there is not a simple relationship between the expression of ethnic stereotypes and their "use" in behavior toward specific ethnic group members. Intrapersonal and situational conditions, under which such "use" may or may not take place, have yet to be specified. Further, there has been little development of assessment procedures for determining whether such "use" takes place. Until considerable research concerning these issues is carried out, discussion of the stereotype-behavior relationship will have to remain at a largely speculative level.

The Present State of Knowledge

In spite of the volume of research on ethnic stereotypes that has been carried out for the last 40 years, there seem to be relatively few "advances in knowledge" that have been made about stereotypes and stereotyping. We know that subjects, when placed in the appropriate situation (i.e., the Katz and Braly paradigm), will produce patterns of trait assignments to ethnic groups, which are similar in differing areas of the United States and, to a lesser extent, throughout the world. The patterns of trait assignments are also fairly consistent over time (Gilbert, 1951; Karlins et al., 1969) although there may have been a general weakening of consensus in recent years.

Although this consensus on attributed traits may be of interest from a sociological or statistical standpoint, it tells us very little about stereotypes or stereotyping. In the first place, the methodological weaknesses of the research paradigm make interpretations of the meaning of the experimental results tenuous at best. For example, the consensus that has been found appears to be, at least in part, a function of the methodology used, as Brown (1965) and Eysenck and Crown (1948) have pointed out. Second, the results that are observed seem to have significance only as a "measure of social sharedness, rather than of stereotypy [Scott, 1965, p. 94]." Findings as to the extent that a given generalization is widespread among a group of people (subjects), even if methodological problems are ignored, do *not* prove of any value in elucidating what a stereotype is, how they are developed, or even how many people *hold* a stereotype. For example, how many subjects actually endorse (agree with) or *use* the generalization in behavior, and how many subjects are just reporting on their knowledge of the traits that persons in one's culture most commonly attribute to the given ethnic group (cf. Karlins et al., 1969, p. 15)?

The magnitude of generalizations has also been largely overlooked. Since most research has concentrated on the assignment of traits that are "most typical" of an ethnic group, we do not know whether the subject believes that a trait which is "typical" is found in 20%, 50%, 90%, or all of the members of an ethnic group.[8] While some theorists have appeared to assume that the generalizations involved in stereotyping are virtually exceptionless, recent research (Brigham, 1969; Mann, 1967) has suggested that exceptionless generalizations are quite rare.

The psychological processes underlying stereotyping have been largely bypassed in the research literature. The mechanisms of selective perception, projection, rationalization, and denial are most often used in descriptions of stereotyping, while stereotyping is tied in with the general personality characteristics of ethnocentrism, authoritarianism (Adorno et al., 1950), and concreteness of thinking (White & Harvey, 1965). But little has been done toward the construction of a unified theory of stereotyping. Research into the possible functional values of stereotyping and its implications for differential success of stereotype-change attempts has received little attention.

For instance it may be that for stereotypes, as well as for generally negative attitudes, the functional bases may differ for different individuals (Katz, 1960). Following the work of Katz and his associates (Katz, McClintock, & Sarnoff, 1957; Katz, Sarnoff, & McClintock, 1956; Stotland, Katz, & Patchen, 1959), it could be predicted that stereotypes based on conformity to social norms would be susceptible to change through "reality oriented" techniques. For stereotypes that serve an ego-defensive function, on the other hand, therapeutic techniques such as catharsis and insight could be expected to be effective as agents of change. Unfortunately, as implied previously, present measures of stereotypes do not yield information concerning the functional characteristics of stereotyping.

Many researchers have proposed (with an almost audible sigh): "Stereotyping is universal among all members of the human race [Bogardus, 1950, p. 289]," but most writers agree that stereotypes are undesirable and should be eradicated. However, there is little agreement concerning just what is the facet of stereotypes that makes them undesirable (e.g., inaccuracy, faulty thought processes, unacceptable basis of acquisition, rigidity, ethnocentrism, etc.). This lack of agreement, of course, is tied in with the similar lack of agreement as to what stereotypes are.

Research concerning the *development* of generalizations about ethnic groups, their *rigidity* in the face of contradictory evidence, and especially the *influence* of these generalizations *on actual intergroup behavior* has been infrequent and hampered by differing conceptualizations of the phenomena under investigation. Such research as has been carried out has failed to identify any clear-cut stereotype-behavior relationship. Finally, while the distinction between "social sharedness" or "cultural consensus" and "personal stereotype" has been recognized by several authors, this distinction has made little impact on the research literature.

Despite this present state of affairs, which would appear to be rather bleak in terms of level of knowledge attained, most researchers seem to share the view that the concepts of stereotype and stereotyping nevertheless can be of considerable value and importance in the understanding of human behavior. Since the present writer shares this view also, it would seem germane to attempt a reconceptualization and reformulation of the concept, both in terms of conceptual clarity and in terms of potential for research.

ETHNIC STEREOTYPES: A MODIFIED CONCEPTUAL FRAMEWORK

There seems to be general agreement among psychologists that a stereotype is a generalization (or belief) about a group; most likely, a generalization concerning the perceived prevalence of a given trait within an ethnic group. The term is also often used in a more general sense to describe a particular *pattern* of traits that is commonly assigned to an ethnic group. To avoid confusion, it might be helpful to refer to the latter case as a general or overall "group stereotype," reserving the term "stereotype" to denote a single trait attribution. Certainly, if one wishes to categorize people as to whether or not they *tend* to stereotype, then attention would be given to patterns of trait attribution that they might endorse.

---

[8] During a pilot study, Brigham (1969) had subjects go through the Katz and Braly paradigm, then go *back* and assign percentages for the five starred "most typical" traits for each ethnic group. Percentages assigned ranged from 10% to 100%, with a mean of 55. In many cases, subjects saw *all five* of the "most typical" traits as being characteristic of *less than half* of the members of the ethnic group.

Although it seems clear that a stereotype is a type of generalization, it seems equally clear that not all generalizations are stereotypes. Surely, "stereotype" is not simply a synonym for "generalization"; the former term must have meaning above and beyond that attributed to the latter. What is it, then, that differentiates a stereotype from an "ordinary" generalization? Or, more simply, what differentiates a "good" generalization from a "bad" generalization (i.e., stereotype)? Since none of the currently available conceptualizations seems completely adequate, it must be the case that some aspect of the use of stereotype as a concept is being overlooked. If attention is given to those aspects of stereotypes that seem objectionable, a more adequate conceptualization might arise.

If there is any single theme running through the views presented in this study, it would seem to be that when a person's belief is described as a stereotype, he is being criticized for having failed *in some way* to meet rational standards of objectivity. The shortcomings of previous treatments of stereotype involve the assumption that the attributed lack of objectivity is always of one type—namely, that of accepting a generalization that is false, or is unsupported by available data, or is based upon illogical reasoning, or is held with inordinate rigidity, and so forth.

The endorsement of a stereotype can be seen as involving a number of different errors. Perhaps the term most adequate for capturing the nature of these errors is *unjustified*. This term can have two distinct senses, both of which may be relevant. On the one hand, as suggested above, a generalization can be seen as unjustified when it does not correspond to the facts, or when one has not gathered sufficient evidence to support it. On the other hand, stereotypes may enter into verbal behavior as justifications for prejudicial or discriminatory social practices, which themselves may be viewed as unjustified. It is this latter use to which our attention has been called by the notion of "white racism" (e.g., Kerner et al., 1968). When practices of discrimination are all-pervasive, justification for such behavior is usually not necessary. However, as Brown, Campbell, and others have pointed out, should these practices be criticized, then such generalizations may be available to enable one to consider the discriminatory practices to be based upon natural, innate, or irreversible intergroup differences.

One facet of the use of this concept which has not been commented on is the fact that the term is usually applied by someone to *someone else's* generalization. That is, one does not typically label his own beliefs as stereotypes; the label is generally affixed by someone else who is observing, interpreting, or describing a person's beliefs. The term stereotype, therefore, is generally attributed to a person's beliefs by someone else who is in some way observing the situation, and the criteria used in classing the beliefs as stereotypes are those of the observer. It is the observer who makes the decision that the beliefs do not meet his own criteria of acceptability and hence are stereotypes. (In a similar manner, certain negative attitudes may be labeled "prejudicial" by an observer of this situation.)

The concept of stereotype may therefore be defined in the following manner: *An ethnic stereotype is a generalization made about an ethnic group, concerning a trait attribution, which is considered to be unjustified by an observer*. It is this observer who will decide that the generalization is or is not a stereotype, according to his criteria of justifiability.

In most of the theorizing and research reviewed herein, these observers have been psychologists or sociologists. Viewed under this frame of reference, then, the major shortcoming of most of the published "stereotype" research (ignoring, for the moment, methodological inadequacies) has been a failure on the part of these observers to make explicit the criteria under which they are considering the generalizations unjustified. And, in those cases where criteria have been described, such as factual incorrectness, representations of inferior judgmental processes, or justifications for discriminatory behavior, it has usually been left unsaid that these criteria remain *assumptions* on the part of the particular researchers, rather than measurable attributes of the situation.[9]

Some writers, for example, have proposed that the high intersubject agreement on the attribution of certain traits to certain ethnic groups provides evidence that these attributions are invalid. Why should this be so? It seems to this writer that the most reasonable interpretation of this approach would be that the researchers (observers) are *assuming* that such modal trait attributions *must* have been acquired from hearsay (or from an inadequate basis of acquisition) rather than through an "acceptable" learning situation. Yet, with a few exceptions (e.g., Brown, 1965; Klineberg, 1951) this latter point has not been made. Even in cases where this approach has been formally stated, furthermore, writers have failed to note that until data bearing on the basis of acquisition are collected, such an interpretation must be regarded as an assumption or a working hypothesis.

The chief point of this reformulation, therefore, is to suggest that these assumptions (of unjustifiableness) *must* be made explicit by those theorists and researchers interested in ethnic stereotypes. If this is done, the opportunities for important and meaningful research into stereotypes, the way in which they are learned, the functions they may serve for people who hold them, their relationship to attitudes, and their importance as determinants of behavior would seem to be considerably increased. Suggestions for some potentially productive paths of research in these areas are outlined below.

Implications for Research

Measurement of stereotypes. Under the Katz and Braly format, a stereotype of an ethnic group has been treated as the traits most commonly attributed to that group by a sample of subjects. However, analysis of individuals and the particular traits that they attribute has seldom been attempted. It appears that most researchers have assumed that the more an individual agrees with the trait attributions made by the entire sample, the more he is giving evidence of stereotyping. The rationale underlying this assumption may be that agreement with modal trait attributions is seen as presumptive evidence that such views have been obtained from an unacceptable source, for instance, hearsay, rather than from face-to-face contact, and are hence unjustified.

However, this assumption has not been tested. The sources from which subjects obtain their ideas about which traits to attribute to different ethnic groups have not been elucidated. Therefore, the implicit assumption that such trait attributions are somehow unjustified would seem to be tenuous at best. Even if this assumption is accepted, furthermore, few investigators have been interested in assessing the degree to which individuals agree with these modal trait attributions, that is, deriving individual stereotyping scores. To this writer's knowledge, only one such study (Brigham, 1969) has been carried out under the Katz and Braly paradigm.

Since this paradigm suffers from so many methodological shortcomings, as pointed out earlier, other methods of assessing subjects' beliefs about the prevalence of traits within ethnic groups would certainly be welcome. One possible method might be the previously discussed technique of having subjects circle the *percentage* of the ethnic group thought to have each of a number of listed traits. This method avoids the ambiguity that seems to accompany the concept of "typical" in the Katz and Braly paradigm, while the researcher's (observer's) conception of justified can be inserted readily. It may be that the researcher feels that any *extreme* percentage assigned (e.g., over 80% or under 20%) is unjustified and hence a stereotype. Or, it could be proposed that any percentage deviating by more than one (or two) standard deviations from the sample mean would be considered unjustified and a stereotype. In either case, each subject can then be given a "stereotyping score" indicating the number and/or intensity of the stereotypes that he endorsed.

---

[9] Another criteria for justification could be the function which the observer sees the stereotype as serving for the subject. If the researcher wishes to class the *function* as unjustified, for instance, due to the "self-delusional" nature of ego-defensive beliefs (Katz, 1960), then it follows that those ethnic trait attributions that are seen as performing that function will be classed as stereotypes.

A modification of this approach would be to ask subjects to give the *range* of percentages which they would think are "reasonable" or "justified" for each trait relevant to the ethnic group. The subjects in this case could be either representatives of the populations under investigation (e.g., college students) or representatives of the "observer population" (e.g., social psychologists). The modal range given by the subjects for each trait could then be used as criteria for the classifying of the responses of other subjects as justified or unjustified; that is, circled percentages that are outside of the obtained modal range for a particular trait would be considered unjustified and hence a stereotype. Again, each subject would then have a stereotyping score, denoting the number of trait attributions for which his circled percentages did not fall within the (modal) range of percentages seen as "justified" by the previous sample.

A further extension of this technique could involve instances in which trait attributions to two groups are assessed. If, in the researcher's (observer's) opinion, these two groups should be regarded equally (for instance, Jews and Christians, or American Negroes and whites), stereotyping could be operationally defined as instances wherein trait attributions were *not* equivalent for the same trait. The magnitude of an individual's stereotyping could, then, be seen as a continuous variable, representing the difference between the percentages he attributed to the two groups for the same trait(s).

Learning of stereotypes. As indicated earlier, very few researchers have devoted their attention to the way in which ethnic generalizations are learned and the sources from which they are learned. Although a few cross-sectional studies of the content of generalizations endorsed by children of different ages have been carried out (e.g., Blake & Dennis, 1943), it would seem that studies making use of longitudinal data are of great importance. In a related vein, an investigation of the sources that are cited, by children of differing ages, as the basis for their ethnic generalizations would be of great interest.

We might expect that as a child grows older and is exposed to an ever-widening sphere of social influence agents, his views and the sources from which he derives them could be considerably modified. It could be hypothesized that correspondence with parental views regarding ethnic trait attributions might be lessened, at least in situations where the information imparted by other sources, for instance, peers, teachers, and communications media, differs from that supplied by the parents. Even with relatively young children, questions of the "How do you know that?" type might provide considerable evidence regarding the source from which ethnic generalizations are learned.

Degree of commitment and saliency. As many attitude theorists have pointed out, research on attitudes has typically ignored the saliency, or importance, of attitudes to individuals. As noted, the conditions under which the attitude is learned, and the conditions which may induce change in an attitude, may differ for attitudes of great and little saliency. Similarly, the function that an attitude may serve for the holder's adjustment (Katz, 1960) and the degree to which it is interrelated with other similar attitudes can have major consequences for the success of attitude-change attempts. The same possibilities exist for the analysis of stereotypes and their possible interrelations with other aspects of personality.

A logical derivation from the research findings on attitudes toward ethnic groups would lead to the hypothesis that those stereotypes which are salient to the individual, to which he is committed (behaviorally, or verbally), that are buttressed by the endorsement of other similar stereotypes, and that are serving an ego-defensive function for the personality, would be least amenable to change by rational, informational methods.

Little or no empirical attention has been paid to the degree of commitment that subjects have toward the generalizations that

TABLE 1

Correlations between Attitude toward Negroes and Traits Attributed to Negroes, Percentages of Negroes Estimated to "Have" Each Trait, and Differences in Percentages Assigned to Negroes and (White) Americans for the Same Trait

Trait	Katz and Braly paradigm		Percentage-assignment paradigm	
	Percentage of subject's attributing	Biserial correlation with attitude	Percentage assigned to Negroes × Attitude	Percentage assigned to Americans minus percentage assigned to Negroes[a] × Attitude
Intelligent	11	.29**	.24***	.32***
Industrious	8	.17	.33***	.32***
Friendly	40.5	.09	.22***	.30***
Nationalistic	16.5	.20*	.03	−.04
Religious	26.5	−.11	.02	.16**
Aggressive	20	.14	−.04	−.03
Talkative	30	.06	−.04	−.04
Musical	59.5	−.08	−.05	[b]
Athletic	71.5	.04	−.08	[b]
Happy-go-lucky	24	−.17*	−.06	[b]
Hostile	27.5	.09	−.13*	−.14**
Impulsive	18.5	−.06	−.22***	−.11
Revengeful	24	−.13	−.22***	[b]
Spend money unwisely	13.5	−.47***	−.36***	[b]
Irresponsible	18.5	−.48***	−.37***	−.40***

Note.—High score denotes a pro-Negro attitude; $N = 200$.

[a] Difference scores were reflected in some cases so that a high score always denotes a pro-Negro orientation, that is, relatively higher percentages circled for positive traits for Negroes and relatively low percentages circled for negative traits.

[b] Traits which were not listed for Americans.

* $p < .07$.

** $p < .05$.

*** $p < .01$ (all significance tests are two-tailed).

they hold. Neither the Katz and Braly paradigm nor the percentage-circling design provides data on the degree to which subjects feel certain that their trait attributions are correct. If some subjects are regarding their responses as wild guesses, while others are fairly certain that their responses are "right" and are valid estimates of reality, then responses of these two types may hold quite different implications for behavior and for the changing of stereotypes. It might be expected, for example, that the use of a weighted score, taking both extremity of stereotype and degree of certainty into account, could improve behavioral prediction considerably.

Stereotype and attitude. While some researchers have found small or moderate relationships between their measures of general stereotyping and of attitude toward the group in question, little attention has been paid to the relationship between ethnic attitudes and *specific* stereotypes (trait attributions). The results of one attempt to get at this relationship are discussed below. Subjects assigned traits to Negroes in the Katz and Braly paradigm and also assigned percentages for 15 traits to "Negroes" and to (presumably, white) "Americans," along with seven other ethnic groups. They were also given a widely validated measure of attitudes toward Negroes (Brigham, Woodmansee, & Cook[10]; Woodmansee & Cook, 1967).

Correlations between the attitude measure and trait attributions, percentages circled, and "difference" percentages circled (Negro minus American) for the same traits are presented in Table 1. As can be seen, attribution of traits in the Katz and Braly framework was significantly related to racial attitude for 3 of the 15 traits in question, while the relationship with racial attitude was significant for 7 of the 15 percentages circled for Negroes, and for 6 of the 10 "difference" percentages. Interestingly, the 2 traits most often attributed to Negroes, and for which the mean percentages circled were highest—Athletic and Musical—showed no relationship to racial attitude. Percentages assigned to Negroes for all 15 traits, when placed in a stepwise regression analysis, yielded a multiple R of .52 with racial attitude.

While this covariance between racial attitudes and beliefs about the prevalence of traits in Negroes is certainly not surprising, the data do provide information about the degree to which the attribution of specific traits covaries with attitude. Of further interest would be the degree to which the overall "favorability" of specific attributed traits is linearly related to attitude. For instance, is the attribution of traits that may be neutral or favorable in affective tone (such as, perhaps, happy-go-lucky, cheerful, sense of humor, pleasure-loving) an indication of a favorable attitude, or are such traits indicative of a negative attitude when attributed to certain ethnic groups, such as Negroes?

---

[10] Brigham, J. C., Woodmansee, J. J., & Cook, S. W. Unpublished manuscript entitled "Dimensions of verbal racial attitudes: Racial intermarriage and Negro self-improvement."

Stereotypes and behavior. As outlined earlier, little research has been carried out relating stereotypes to actual behavior toward members of the object group. While attitudes have shown a small to moderate relationship with behavior toward group members in some situations (e.g., Warner & DeFleur, 1969), it is an open question as to whether knowledge of a person's stereotypes allows any increase in predictability beyond that yielded by measures of general affect, that is, attitude. One research attempt (Brigham, 1969) suggests that knowledge of subjects' stereotypes of Negroes adds little to predictive power beyond that attributable to general racial attitude.

In the two studies cited wherein evaluation of specific object group members was assessed with reference to those traits included in individuals' stereotypes (Brigham, 1969; Mann, 1967a) the results have suggested that specific generalizations may *not* carry over into evaluations of ethnic group members on that particular trait. Instead, many subjects evaluated ethnic group members in a manner inconsistent with the generalizations they had endorsed about the ethnic group. The implications of such findings are not clear.

As stated near the beginning of this paper, this writer shares the conviction that ethnic stereotypes may serve as significant determinants of behavior toward ethnic group members. Yet, on the basis of data available thus far, such a conviction must remain only a hypothesis. It seems at least possible that much of the ambiguity in this area can be resolved with the development of more sophisticated instruments for assessing stereotypes. If instruments are developed that can measure saliency of the stereotype for the subject, and the certainty with which he holds it, for example, predictive power may increase. Naturally, consideration must also be given to the situational and personality variables that bear on any behavioral situation. With the development of better assessment procedures and the ability to control or at least specify the effects of various situational and personality factors, the role of ethnic stereotypes as determinants of evaluations of, and behavior toward, ethnic group members can be better elucidated.

REFERENCES

Abate, M., & Berrien, F. K. Validation of stereotypes: Japanese versus American students. *Journal of Personality and Social Psychology,* 1967, **7,** 435–438.

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., & Sanford, R. N. *The authoritarian personality.* New York: Harper and Row, 1950.

Allport, G. W. *The nature of prejudice.* Reading, Mass.: Addison-Wesley, 1954.

Allport, G. W., & Kramer, B. M. Some roots of prejudice. *Journal of Psychology,* 1946, **22,** 9–39.

Asher, S. R., & Allen, V. L. Racial preference and social comparison processes. *Journal of Social Issues,* 1969, **25,** 157–166.

Bastide, R., & van Den Berghe, P. Stereotypes, norms, and interracial behavior in Sao Paulo, Brazil. *American Sociological Review,* 1957, **22,** 688–694.

Bayton, J. A. The racial stereotypes of Negro college students. *Journal of Abnormal and Social Psychology,* 1941, **36,** 97–102.

Bayton, J. A., McAlister, L. B., & Hamer, J. Race-class stereotypes. *Journal of Negro Education,* 1956, **25,** 75–78.

Blake, T., & Dennis, W. The development of stereotypes concerning the Negro. *Journal of Abnormal and Social Psychology,* 1943, **38,** 525–531.

Bogardus, E. S. Stereotypes versus sociotypes. *Sociology and Social Research,* 1950, **34,** 286–291.

Brigham, J. C. *Ethnic stereotypes, attitudes, and treatment of ethnic group members.* (Doctoral dissertation, University of Colorado) Ann Arbor, Mich.: University Microfilms, 1969. No. 70–5822.

Brown, R. *Words and things.* Glencoe, Ill.: The Free Press, 1958.

Brown, R. *Social psychology.* New York: The Free Press, 1965.

Bruner, J. S., & Perlmutter, H. W. Compatriot and foreigner: A study of impression formation in three countries. *Journal of Abnormal and Social Psychology,* 1957, **55,** 253–260.

Buchanan, W. Stereotypes and tensions as revealed by the UNESCO international poll. *International Social Science Bulletin,* 1951, **3,** 515–528.

Campbell, D. T. Stereotypes and the perception of group differences. *American Psychologist,* 1967, **22,** 817–829.

Centers, R. An effective classroom demonstration of stereotypes. *Journal of Social Psychology,* 1951, **34,** 41–46.

Chandra, S. Stereotypes of university students toward different ethnic groups. *Journal of Social Psychology,* 1967, **71,** 87–94.

Child, I. L., & Doob, L. W. Factors determining national stereotypes. *Journal of Social Psychology,* 1943, **17,** 203–219.

Clark, K. B., & Clark, Y. K. The development of consciousness of self and the emergence of racial identification in Negro pre-school children. *Journal of Social Psychology,* 1939, **10,** 591–599.

Das, J. P. Ethnocentrism and stereotypes among Santhals. *Journal of Social Psychology*, 1962, **57**, 79–84.

DeBie, P. Certain psychological aspects of Benelux. *International Social Science Bulletin*, 1951, **3**, 540–552.

Dewey, R., & Humber, W. J. *An introduction to social psychology*. New York: Macmillan, 1966.

Diab, L. N. National stereotypes and the "reference group" concept. *Journal of Social Psychology*, 1962, **57**, 339–351.

Diab, L. N. Factors affecting studies of national stereotypes. *Journal of Social Psychology*, 1963, **59**, 29–40. (a)

Diab, L. N. Factors determining group stereotypes. *Journal of Social Psychology*, 1963, **71**, 3–10. (b)

Dudycha, G. J. The attitudes of college students toward war and the Germans before and during the Second World War. *Journal of Social Psychology*, 1942, **15**, 317–324.

Edwards, A. L. Four dimensions in political stereotypes. *Journal of Abnormal and Social Psychology*, 1940, **35**, 566–572.

Ehrlich, H. J. Stereotyping and Negro-Jewish stereotypes. *Social Forces*, 1962, **41**, 171–176. (a)

Ehrlich, H. J. The study of prejudice in American social science. *Journal of Intergroup Relations*, 1962, **3**, 117–125. (b)

Ehrlich, H. J., & Rinehart, J. W. A brief report on the methodology of stereotype research. *Social Forces*, 1965, **43**, 564–575.

Elliott, D. N., & Wittenberg, B. H. Accuracy of identification of Jewish and non-Jewish photographs. *Journal of Abnormal and Social Psychology*, 1955, **51**, 339–341.

Eysenck, H. J., & Crown, S. National stereotypes: An experimental and methodological study. *International Journal of Opinion and Attitude Research*, 1948, **2**, 26–39.

Fischer, H., & Trier, U. P. *Das Verhaeltnis zwischen Deutchsweizer und Westschweizer: Eine sozialpsychologische untersuchung*. Bern: Hans Huber, 1962. Cited by H. C. Triandis & V. Vassiliou. Frequency of contact and stereotyping. *Journal of Personality and Social Psychology*, 1967, **7**, 316–328.

Fishman, J. A. An examination of the process and function of social stereotyping. *Journal of Social Psychology*, 1956, **43**, 27–64.

Gilbert, G. M. Stereotype persistence and change among college students. *Journal of Abnormal and Social Psychology*, 1951, **46**, 245–254.

Goodman, M. E. *Race awareness in young children*. Cambridge, Mass.: Addison-Wesley, 1952.

Greenwald, H. J., & Oppenheim, D. B. Reported magnitude of self-misidentification among Negro children—Artifact? *Journal of Personality and Social Psychology*, 1968, **8**, 49–52.

Harding, J., Kutner, B., Proshansky, H., & Chein, I. Prejudice and ethnic relations. In G. Lindzey (Ed.), *Handbook of social pschology*. Vol. II. Cambridge, Mass.: Addison-Wesley, 1954.

Harding, J., Proshansky, H., Kutner, B., & Chein, I. Prejudice and ethnic relations. In G. Lindzey and E. Aronson (Eds.), *Handbook of social psychology*. Vol. V. Cambridge, Mass.: Addison-Wesley, 1969.

Himmelfarb, S. Factors related to the perception of "Jewishness." *American Psychologist*, 1960, **15**, 401. (Abstract)

Hoult, T. F. An effective classroom demonstration of stereotypes: A reexamination. *Journal of Social Psychology*, 1953, **38**, 293–295.

Jahoda, G. Nationality preferences and national stereotypes in Ghana before independence. *Journal of Social Psychology*, 1959, **50**, 165–174.

Janis, I. L., & Smith, M. B. Effects of education and persuasion on national and international images. In H. C. Kelman (Ed.), *International behavior: A social-psychological analysis*. New York: Holt, Rinehart and Winston, 1965.

Karlins, M., Coffman, T. L., & Walters, G. On the fading of social stereotypes: Studies in three generations of college students. *Journal of Personality and Social Psychology*, 1969, **13**, 1–16.

Katz, D. The functional approach to the study of attitudes. *Public Opinion Quarterly*, 1960, **24**, 163–204.

Katz, D., & Braly, K. Racial stereotypes in one hundred college students. *Journal of Abnormal and Social Psychology*, 1933, **28**, 280–290.

Katz, D., & Braly, K. Racial prejudice and racial stereotypes. *Journal of Abnormal and Social Psychology*, 1935, **30**, 175–193.

Katz, D., McClintock, C., & Sarnoff, D. The measurement of ego-defense as related to attitude change. *Journal of Personality*, 1957, **25**, 465–474.

Katz, D., Sarnoff, D., & McClintock, C. Ego-defense and attitude change. *Human Relations*, 1956, **9**, 27–45.

Katz, D., & Stotland, E. A preliminary statement of a theory of attitude structure and change. In S. Koch (Ed.), *Psychology: A study of science*. Vol. III. New York: McGraw-Hill, 1959.

Kerner, O., Lindsay, J. V., Harris, F. R., Brooke, E. W., Corman, J. C., McCulloch, W. M., Abel, I. W., Thornton, C. B., Wilkins, R., Peden, K. G., & Jenkins, H. *Report of the National Advisory Commission on Civil Disorders*. New York: Bantam Books, 1968.

Kerr, M. An experimental investigation of national stereotypes. *Sociological Review*, 1943, **35**, 37–43.

Klineberg, O. The scientific study of national stereotypes. *International Social Science Bulletin*, 1951, **3**, 505–515.

Klineberg, O. *Social psychology*. (Rev. ed.) New York: Henry Holt, 1954.

Koenig, F. W., & King, M. B. Cognitive simplicity and prejudice. *Social Forces*, 1962, **40**, 220–222.

Koenig, F. W., & King, M. B. Cognitive simplicity and out-group stereotyping. *Social Forces*, 1964, **42**, 324–327.

Krech, D., Crutchfield, R. S., & Ballachey, E. L. *Individual in society*. New York: McGraw-Hill, 1962.

LAPIERE, R. T. Type-rationalizations of group antiplay. *Social Forces,* 1936, **15**, 232–237.

LAVIOLETTE, F., & SILVERT, K. H. A theory of stereotypes. *Social Forces,* 1951, **29**, 237–257.

LEVINE, R. A. Socialization, social structure, and intersocietal images. In H. C. Kelman (Ed.), *International behavior: A social psychological analysis.* New York: Holt, Rinehart and Winston, 1965.

LINDZEY, G., & ROGOLSKY, S. Prejudice and identification of minority group membership. *Journal of Abnormal and Social Psychology,* 1950, **45**, 37–53.

LIPPMANN, W. *Public opinion.* New York: Harcourt, Brace, 1922.

LITTERER, O. F. Stereotypes. *Journal of Social Psychology,* 1933, **4**, 59–69.

MANN, J. W. Inconsistent thinking about group and individual. *Journal of Social Psychology,* 1967, **71**, 235–245. (a)

MANN, J. W. Inconsistent impressions in assessing individual before group. *Psychologia Africana,* 1967, **2**, 143–150. (b)

MEAD, M. The cross-cultural approach to the study of personality. In J. L. McCary (Ed.), *Psychology of personality.* New York: Grove Press, 1956.

MEENES, M. A comparison of racial stereotypes of 1935 and 1942. *Journal of Social Psychology,* 1943, **17**, 327–336.

MILGRAM, S. Nationality and conformity. *Scientific American,* 1961, **205**, 45–51.

MORLAND, J. K. Racial acceptance and preference of nursery school children in a southern city. *Merrill Palmer Quarterly,* 1962, **8**, 271–280.

PERLOE, S. I. The effects of attitudes and concepts on categorization. *Journal of Personality,* 1964, **32**, 249–269.

PROTHRO, E. T., & MELIKIAN, L. H. Studies in stereotypes: III. Arab students in the Near East. *Journal of Social Psychology,* 1954, **40**, 237–243.

PROTHRO, E. T., & MELIKIAN, L. H. Studies in stereotypes: V. Familiarity and the kernel of truth hypothesis. *Journal of Social Psychology,* 1955, **41**, 3–10.

RICE, S. A. "Stereotypes": A source of error in judging human character. *Journal of Personnel Research,* 1926, **5**, 267–271.

RIFFENBURGH, R. H. A method of sociometric identification on the basis of multiple measurements. *Sociometry,* 1966, **29**, 280–290.

ROKEACH, M. *The open and closed mind.* New York: Basic Books, 1960.

ROKEACH, M. *Beliefs, attitudes, and values.* San Francisco: Jossey-Bass, Inc., 1968.

SAENGER, G. *The social psychology of prejudice.* New York: Harper, 1953.

SAENGER, G., & FLOWERMAN, S. Stereotypes and prejudicial attitudes. *Human Relations,* 1954, **7**, 217–238.

SANFORD, N. The approach of the authoritarian personality. In J. L. McCary (Ed.), *Psychology of personality.* New York: Grove Press, 1956.

SCHOENFELD, N. An experimental study of some problems relating to stereotypes. *Archives of Psychology,* 1942, **38**, No. 270.

SCHUMAN, H. Social change and the validity of regional stereotypes in East Pakistan. *Sociometry,* 1966, **29**, 428–440.

SCODEL, A., & AUSTRIN, H. The perception of Jewish photographs by non-Jews and Jews. *Journal of Abnormal and Social Psychology,* 1957, **54**, 278–280.

SCOTT, W. A. Psychological and social correlates of international images. In H. C. Kelman (Ed.), *International behavior: A social psychological analysis.* New York: Holt, Rinehart and Winston, 1965.

SECORD, P. F. Stereotyping and favorableness in the perception of Negro faces. *Journal of Abnormal and Social Psychology,* 1959, **59**, 309–315.

SECORD, P. F., & BACKMAN, C. W. *Social psychology.* New York: McGraw-Hill, 1964.

SECORD, P. F., BEVAN, W., JR., & KATZ, B. The Negro stereotype and perceptual accentuation. *Journal of Abnormal and Social Psychology,* 1956, **53**, 78–83.

SHERIF, M., & SHERIF, C. W. *Social psychology.* New York: Harper and Row, 1969.

SIMPSON, G. E., & YINGER, J. M. *Racial and cultural minorities.* (Rev. ed.) New York: Harper and Row, 1965.

SINHA, A. K. P., & UPADHYAY, O. P. Stereotypes of male and female university students in India toward different ethnic groups. *Journal of Social Psychology,* 1960, **51**, 93–102. (a)

SINHA, A. K. P., & UPADHYAY, O. P. Change and persistence in the stereotypes of university students toward different ethnic groups during Sino-Indian border dispute. *Journal of Social Psychology,* 1960, **52**, 31–39. (b)

STOTLAND, E., KATZ, D., & PATCHEN, M. The reduction of prejudice through the arousal of self-insight. *Journal of Personality,* 1959, **27**, 507–531.

TAFT, R. Ethnic stereotypes, attitudes, and familiarity: Australia. *Journal of Social Psychology,* 1959, **49**, 177–186.

TAYLOR, R. G., JR. Racial stereotypes in young children. *Journal of Psychology,* 1966, **64**, 137–142.

TRIANDIS, H. C. Interpersonal relations in international organizations. *Organizational Behavior and Human Performance,* 1967, **1**, 26–55.

TRIANDIS, H. C., & VASSILIOU, V. Frequency of contact and stereotyping. *Journal of Personality and Social Psychology,* 1967, **7**, 316–328.

VAN DEN BERGHE, P. L. Race attitudes in Durban, South Africa. *Journal of Social Psychology,* 1962, **57**, 55–72.

VINACKE, W. E. Stereotyping among national-racial groups in Hawaii: A study in ethnocentrism. *Journal of Social Psychology,* 1949, **30**, 265–291.

VINACKE, W. E. Explorations in the dynamic process of stereotyping. *Journal of Social Psychology,* 1956, **43**, 105–132.

VINACKE, W. E. Stereotypes as social concepts. *Journal of Social Psychology,* 1957, **46**, 229–243.

WARNER, C. G., & DEFLEUR, M. L. Attitude as an interactional concept: Social constraint and social distance as intervening variables between attitudes and action. *American Sociological Review,* 1969, **34**, 153–169.

WEDGE, B. Nationality and social perception. *Journal of Communication,* 1966, **16**, 273–282.

WELLS, W. D., GOI, F., & SEADER, S. A change in a product image. *Journal of Applied Psychology,* 1958, **42**, 120–121.

WHITE, B. J., & HARVEY, O. J. Effects of personality on own stand on judgment and production of statements about a central issue. *Journal of Experimental Social Psychology,* 1965, **1**, 334–347.

WOODMANSEE, J. J., & COOK, S. W. Dimensions of verbal racial attitudes: Their identification and measurement. *Journal of Personality and Social Psychology,* 1967, **7**, 240–250.

ZAIDI, S. M., & AHMED, M. National stereotypes of university students in East Pakistan. *Journal of Social Psychology,* 1958, **47**, 387–395.

ZAWADZKI, B. Limitations in the scapegoat theory of prejudice. *Journal of Abnormal and Social Psychology,* 1948, **43**, 127–141.

(Received January 8, 1970)

Manuscripts Accepted for Publication in the

Psychological Bulletin

To Dispel Fantasies about Fantasy-Based Measures of Achievement Motivation: Doris R. Entwistle*: Department of Social Relations, Johns Hopkins University, Baltimore, Maryland 21218.

Birth Order Effects: Not Here, Not Now!: Carmi Schooler*: National Institute of Mental Health, Laboratory of Socio-Environmental Studies, Building 10, Room 3D41, Bethesda, Maryland 20014.

The 47, XYY Male: A Review: David R. Owen*: Department of Psychology, University of Iowa, Iowa City, Iowa 52240.

Malnutrition and Mental Deficiency: B. J. Waitzkin*: Department of Psychology, Brandeis University, Waltham, Massachusetts 02154.

---

* Asterisk indicates author for whom the address is supplied.

Psychological Bulletin
1971, Vol. 76, No. 1, 39–44

MEASURES OF CATEGORY CLUSTERING IN FREE RECALL [1]

FREDERICK FRANKEL AND MICHAEL COLE [2]

The Rockefeller University

A major problem in the study of organization in memory has been the failure to agree upon an adequate measure of category clustering. A *Z*-score measure, based on a test of the number of runs of similar items in the recall list, is proposed. Several alternative measures, including some that are widely used in the literature, are reviewed and their inadequacies discussed. A comparison between the various measures is made using data obtained from a simulated experiment and a previously published developmental study. It is concluded: (*a*) The measures differ widely in the magnitude of effects they yield; (*b*) alternative measures tend to accept the null hypothesis when the *Z* score indicates a significant difference; (*c*) certain measures are biased to increase solely as a function of longer recall list lengths.

Since its inception, the study of organization in the free recall paradigm has sought to account for the fact that the amount of clustering usually increases with the number of items recalled (Bousfield, 1953; Shuell, 1969).

A recurring problem in this field has been failure to agree on an adequate measure of clustering. The current review proposes a more adequate measure of clustering than any suggested thus far, and demonstrates the shortcomings of other measures.

There are certain minimal requirements that we have sought in a measure for the degree of category clustering:

1. The measure should be independent of any phenomena that it may be used to explain or predict. In particular, the clustering measure must not increase simply as a function of the number of items in the recall list.

2. The measure should adequately account for chance. First, the constitution of the recall list should be considered: Items from a given category are more likely to be recalled successively if there is only 1 item from another category than if there are 10 items from other categories appearing on the recall list. Second, statistical reliability should increase with increasing numbers of items recalled; maximal clustering for a large recall list should be given more weight than that for a smaller list.

3. The clustering score should mean the same thing, regardless of the parametric values that yielded it. For example, a difference score has been devised in which the degree of clustering expected by chance is subtracted from the observed amount of clustering on a recall list (see below). The question is, Does a difference score of (say) 2 have an unambiguous meaning even though it may result from an expected value of 7 and an observed value of 9, or an expected value of 2 and an observed value of 4?

The Runs Test as a Measure of Clustering

A convenient way to represent the degree of clustering in a recall list, which is in accord with these principles, is to express clustering in terms of the number of runs of items from a given category present in the recall list. The length of a run is defined as the number of successively recalled items from the same category, regarding single items as runs of one. The number of runs is one more than the number of times during recall that the category of the items being recalled changes. For example, suppose that there are four categories recalled in a given trial: *a*, *b*, *c*, and *d*. The recall list *aaabcdcc* would have a total of five runs starting with a run of three and ending with a run of two.

Mood (1940) has provided the statistics for the mean, M_r, and variance, V_r, for the number of runs in a randomly selected list of arbitrary length N and category composition N_j. Wallis and Roberts (1957, p. 571) have adapted the formulas to a form useful for the present problem:

---

[1] Preparation of this paper was supported in part by National Science Foundation Grant G1221 to M. Cole and in part by Public Health Service Grant GM16735 from National Institute of General Medical Sciences. The ideas on which this paper is based grew out of discussions with A. K. Romney and I. Birnbaum to whom we give thanks for pointing out the general applicability of the runs test and certain inadequacies of alternatives.

[2] Requests for reprints should be sent to M. Cole, The Rockefeller University, New York, New York 10021.

$$M_r = \frac{N(N+1) - \sum_j N^2}{N}, \qquad [1]$$

and

$$V_r = \frac{\sum_j N_j^2[\sum_j N_j^2 + N(N+1)] - 2N\sum_j N_j^3 - N^3}{N^2(N-1)}, \qquad [2]$$

for a recalled list of N items, with N_j items present from the jth category.

The mean number of runs, M_r, is normally distributed so that knowing M_r, V_r, and the observed number of runs (O_r), a Z score can be calculated:

$$Z = \frac{O_r - M_r}{\sqrt{V_r}}$$

and reference can be made to a normal distribution table for the approximate probability that any individual clustering Z score is a chance occurrence.[3] We define clustering as the presence of significantly "too few" runs. A significant *positive* Z score is the convention adopted here for simplicity.

This procedure is easily adapted to calculation of the degree of clustering observed in a group's recall. Because we are operating with Z scores, this formula reduces to $\sqrt{n}\cdot\bar{Z}$, where n is the number of subjects in the group and $\bar{Z}$ is the mean Z score for the group.

This measure has several advantageous properties. First, it takes account of the categorical composition of the recall list: The same observed number of runs and number recalled for two recall lists will yield different Z scores if the numbers of items recalled from the different categories (N_js) differ. For example, again using the letters a, b, c, and d to represent different members of the same categories, the recall list *aaabbcac* yields a Z score of .94 (1.00/1.07). (No repetitions or intrusions are assumed. A simple generalization of the technique used here can be applied to lists in which items are repeated or intruded.) Another recall list, *aaabbaba*, having the same N and O_r, yields a Z score of $-$ .21 ($-$ .25/1.2).

Second, perfect clustering is given higher scores with higher N: the recall list *aaaaabbbbbcccccddddd* yields a $Z = 6.75$; the list *aaabbbccdd* yields a $Z = 3.92$; and the list *aaaabbbb* yields a $Z = 2.29$. All three lists are perfectly clustered. The Z-score differences arise from a statistical consideration: V_r (Formula 2), having $N(N-1)$ in the denominator, decreases with increasing N.

Third, values generated by different Ns and N_js are comparable to one another because they are obtained in relation to the same, normal distribution, thereby removing any length-related bias. As the following examples illustrate, the relation between degree of clustering and N is by no means intuitive (note that any recall list having a Z score greater than 1.96 reflects statistically significant clustering, $p < .05$): *abccdd* ($Z = 1.91$), *abbccdd* ($Z = 2.03$), and *aabccccbddcd* ($Z = 1.97$).

We will now show that all of these properties are not present in any one of several alternative measures of clustering.

Alternative Measures of Clustering

Several alternative approaches to the measurement of clustering have appeared in the literature. One approach has been to express clustering in terms of the number of pairs of successive items from the same category (NP) that occur during recall.

The number of pairs within any given run has to be one less than the length of that run. Comparison of these measures is facilitated by the following notation:

$$NP = \sum_i (m_i - 1) = \sum_i m_i - \sum_i 1, \qquad [3]$$

---

[3] The assumption of true normality requires large N_js. Wallis and Roberts asserted that normality is a reasonable assumption for relatively small N_js, but the distribution present for very small N_js has not been systematically studied. Simulated data reported below make the assumption appear reasonable for the kinds of data usually encountered in recall studies.

where m_i is the number of items in the ith run, and k is the number of runs. Noting that $\sum_i m_i = N$, $\sum_i 1 = k$, and substituting into Formula 3, it can be seen that the relation of the number of runs to the number of pairs is given by the formula

$$k = N - NP. \qquad [4]$$

Ratios involving the number of pairs as a measure of clustering are quite widely used (Cohen & Bousfield, 1956; Lambert, Ignatow, & Krauthamer, 1968). A score in terms of these measures represents the number of pairs of clustered items per word recalled ($LR = NP/N$) in Lambert's measure and per word minus 1 ($RR = NP/N - 1$) in Bousfield's measure. Another similar approach is to count the total number of items appearing in runs larger than 1 and divide by N, yielding a percentage in clusters (*PC*) measure.

Shuell (1969) has suggested that this division by N (or $N - 1$) serves to make the measure independent of N by "partialling it out."

All three of these measures share one serious shortcoming: They fail to account for the categorical constitution of the recall list, and thus the chance likelihood of a given score becomes indeterminate. To take a simple case, if a subject recalls four items with only one pair of clustered items, he would receive scores of .33 on *RR*, .25 on *LR*, and .50 on *PC*. However, if three of the items were from the same category, the event (one pair) is much more likely than if the items were from three different categories, where a maximum of one pair is possible.

This shortcoming has been recognized by several authors. It has been dealt with by calculating the expected number of pairs (E) for a list of N_j items recalled from each of j categories (Bousfield & Bousfield, 1966). E is given by the formula[4]

$$E = \frac{\sum_j N_j^2 - 1}{N}. \qquad [5]$$

Bousfield and Bousfield (1966) suggested construction of a difference score (*DS*), observed minus expected number of pairs ($NP - E$), as a new measure of clustering. This measure has been widely used (e.g., Dallet, 1964; Scott, 1967). As mentioned earlier, a given *DS* is ambiguous, since it may arise from several nonequivalent sources. Also, for short lists the difference between NP and E is of necessity small, while for longer lists it can be quite large. Thus as Shuell (1969) pointed out, mediocre clustering of a large list may yield a larger *DS* than perfect clustering of a smaller list. Note that a larger *DS* given to a list with larger N does not arise from the statistical consideration mentioned earlier.

After reviewing some of the inadequacies of *RR* and *DS*, Shuell (1969) suggested the use of the maximum possible number of pairs (max) for a given set of N_js, in combination with *DS*. Maximum clustering occurs when the number of runs is equal to the number of categories recalled (i.e., all the members of the jth category are recalled successively). Thus, if N_j is the number of items recalled from the jth category, then (substituting into Formula 3):

$$\max = \sum_j (N_j - 1). \qquad [6]$$

Relevant to Shuell's suggestion is a measure devised by Gerjuoy and Spitz (1966), using *DS* in the numerator and max $- E$ in the denominator [$GR = DS/(\max - E)$]. However, *GR* has its own shortcomings: Negative *GR*s are possible, and a *GR* of $-.60$ is not necessarily analogous to a *GR* of .60 because max $- E$ does not necessarily represent the maximum possible negative difference (i.e., the range of this measure is not always -1.00 to 1.00 as one might expect). For instance, with N_js of 3, 1, 1, and $NP = 0$, *GR* yields a -1.5.[5] Roenker, Thompson, and Brown (1971) have pointed out that the range of *GR* is invariant with list size. Although at first this invariance might seem a virtue, it means that the measure does not take into account the increased reliability of observations made on longer recall lists.

One approach has not yet been considered which might be conceived as similar to the Z

---

[4] The expected value of the number of pairs, E, can also be obtained by substituting Formula 4 into Formula 1 which was the expected number of runs, M_r.

[5] In practice, we have found that the obtained distribution of scores for *GR* from actual experiments is extremely negatively skewed (e.g., when applied to the data of Cole et al., 1970, Experiment II), violating the normality assumption of analysis of variance.

score. Bousfield (1953) has suggested the Index of Repetition, based on the normal approximation to the binomial distribution. This measure has rarely been used, even by Bousfield. As Shuell (1969) pointed out, use of the measure requires two controversial assumptions which few investigators are willing to accept: First, the whole list of items is assumed to be available to a subject whose recall contains only a subset of the total. Second, it is assumed that items recalled are sampled from the whole list with replacement.

Since these assumptions do not agree with existing data, this measure is not considered further.

Comparison of the Various Measures

Shuell (1969) calculated correlations between several different measures of clustering using data from an experiment on college students. The correlations obtained were very high, thus giving the impression that the measures compared were generally in good agreement with one another. However, the subject populations were relatively uniform, consisting only of college students, and only within-trial comparisons were made. Both of these factors limit the variability in the number of items recalled and list composition, because this particular subject population tends to organize their recall to a large degree. As is shown, the number recalled and the composition of the recall list independently affect the comparability of the various measures. Thus, Shuell's comparisons could be expected to be insensitive to real differences. In order to provide a more sensitive test of the comparability of various clustering measures, two different methods of comparison were used.

For the purpose of determining the relationship between the magnitude of the clustering measures and increasing N, an artificial experiment was performed using four groups of 200 statistical subjects each. The four groups were constructed so that they differed systematically only in the *number* "recalled"; the particular items "recalled" were selected by a random process from a list of 20 items arbitrarily divided into four categories of 5 items each. The groups were assigned recall lists of 5, 10, 15, and 20 items, respectively, which were chosen at random from the pool of 20. Clustering scores for each of the various measures discussed in the previous section were calculated for all subjects, and analyses of variance were performed. The group means for two of the measures, LR and PC, were found to increase significantly with increasing N ($F = 5.70$ and 4.70, respectively, $df = 3/597$, $p < .01$). No other measure was found significant ($F < 1$).[6]

Why doesn't RR also show this bias? The answer lies in the fact that for RR, 1 is subtracted from the number recalled in the denominator. This has the effect of increasing the clustering scores of recall lists with lower Ns proportionally more than scores of recall lists with higher Ns (the difference between N and $N - 1$ is relatively larger for smaller Ns).

In order to gauge the effects of list composition on the comparability between the various measures, each kind of score was calculated using data obtained in an experiment previously conducted by the authors (Cole, Frankel, & Sharp, 1971). This experiment was especially useful for present purposes, because a relatively wide variety of experimental manipulations and subject populations was involved, providing a good range in variation of both N and composition. Briefly, the variables involved were grade in school, order of stimulus presentation, and type of stimuli. Five trials were given to each subject.

The groups differed significantly with respect to N for all main effects, and with respect to the Z score for all main effects and four of the interactions.

The results of analysis of variance performed on the other measures, in terms of the relative amount of difference between conditions estimated by each of the measures, are presented in Table 1.

Several conclusions can be drawn from comparisons of the proportions in Table 1. When the manipulations produced large differences in the number recalled (N) as in the Grade Effect, all measures yielded similar conclusions. However, the relative amount of an increase in clustering as a function of Grade, estimated by

---

[6] We also noted that RR had a relatively high chance mean of .21.

TABLE 1

PERCENTAGE INCREASE IN THE MEANS OBTAINED FROM THE CLUSTERING MEASURES FOR THE DEVELOPMENTAL STUDY FOR SELECTED EFFECTS

Effect	*PR*	*LR*	*PC*	*DS*	*GR*	*Z* score	*N*
Grade (G)	53.7**	63.0**	37.4**	241.5**	86.0**	27.1**	67.4**
Stimulus order (So)	−26.9**	−26.7**	−23.4**	−38.8**	−43.7**	−39.6**	−9.5**
Stimulus (S)	29.5**	32.8**	23.0**	97.6**	49.5**	79.6**	28.6**
Trials	21.6**	26.6**	11.2**	116.6**	17.0**	73.9**	43.5**
G × S	150.7*	124.4	36.2	164.7**	38.4**	106.5*	33.1
So × S	−46.6**	−45.4*	−48.0**	−41.2*	−48.6	12.6*	−61.2

* $p < .05$.
** $p < .01$.

each measure, varies radically: *DS* estimates a difference 10 times greater than that estimated by the *Z* score.

When effects were of smaller magnitude, and the corresponding *F* ratios smaller, the conclusions reached concerning the various clustering measures differed greatly. For instance, *RR* and the *Z* score yielded *p*s less than .05 for the Grade *x* type-of-stimulus interaction, which would lead us to conclude that the magnitude of this variable was greater for the higher Grades. The effect would not be found significant if *LR*, *PC*, or *GR* were used ($F = 2.96$, .51, and .46, respectively) and would be overestimated if *DS* were used ($F = 6.56, p < .01$).

In general, the alternative measures accepted the null hypothesis in marginal cases when the *Z* score found it to be false.

Summary and Discussion

This review of measures of categorical clustering indicates that standard measures are misleading for a variety of reasons. *RR*, *LR*, and *PC* fail to take account of the categorical makeup of a given recall list, an important determinant of the chance likelihood of obtaining a specified run length. *PC* fails to account for lesser likelihood of larger runs; *LR* and *PC* were found to be biased to increase with increasing numbers of items recalled. Perhaps with larger list lengths (as those more commonly used), *RR* too would be found to be biased.

The earlier solutions designed to account for the chance likelihood of a given clustering score based on *E*, the expected number of pairs in a list of arbitrary composition, were found inadequate for the present problem. The difference score derived from *E*, like the scores which fail to consider chance levels of clustering, has different meaning when obtained from recall list of varying lengths and composition leading to overestimates for longer list lengths and certain list compositions. Inclusion of intuitive normalizing operations such as max − *E* (as in *GR*) in the calculation of the difference score is at best a rough solution, since max − *E* in the denominator is only loosely related to the theoretically appropriate denominator, and does not take statistical reliability into account.

By applying the theory of runs to the present problem it was shown that a *Z*-score measure can be derived which is not subject to the difficulties inherent in other approaches and which has been shown empirically to be more sensitive than the other alternatives discussed.

REFERENCES

BOUSFIELD, A. K., & BOUSFIELD, W. A. Measurement of clustering and sequential constancies in repeated free recall. *Psychological Reports*, 1966, **19**, 935–942.

BOUSFIELD, W. A. The occurrence of clustering in the recall of randomly arranged associates. *Journal of General Psychology*, 1953, **49**, 229–240.

COHEN, B. H., & BOUSFIELD, W. A. The effects of a dual-level stimulus-word list on the occurrence of clustering in recall. *Journal of General Psychology*, 1956, **55**, 51–58.

COLE, M., FRANKEL, F., & SHARP, D. The development of free recall learning in children. *Developmental Psychology*, 1971, **4**, 109–123.

DALLET, K. M. Number of categories and category information in free recall. *Journal of Experimental Psychology*, 1964, **68**, 1–12.

Gerjuoy, I. R., & Spitz, H. H. Associative clustering in free recall: Intellectual and developmental variables. *American Journal of Mental Deficiency*, 1966, **70**, 918–927.

Lambert, W. E., Ignatow, M., & Krauthamer, M. Bilingual organization in free recall. *Journal of Verbal Learning and Verbal Behavior*, 1968, **7**, 207–214.

Mood, A. M. The distribution theory of runs. *Annals of Mathematical Statistics*, 1940, **11**, 367–392.

Roenker, D. L., Thompson, C. P., & Brown, S. C. Comparison of measures for the estimation of clustering in free-recall. *Psychological Bulletin*, 1971, **76**, 45–48.

Scott, K. G. Clustering with perceptual and symbolic stimuli in free recall. *Journal of Verbal Learning and Verbal Behavior*, 1967, **6**, 864–866.

Shuell, T. J. Clustering and organization in free recall. *Psychological Bulletin*, 1969, **72**, 353–374.

Wallis, W. A., & Roberts, H. V. *Statistics: A new approach*. New York: Free Press, 1957.

(Received February 10, 1970)

Psychological Bulletin
1971, Vol. 76, No. 1, 45–48

COMPARISON OF MEASURES FOR THE ESTIMATION OF CLUSTERING IN FREE RECALL[1]

DANIEL L. ROENKER, CHARLES P. THOMPSON,[2] AND SAM C. BROWN
Kansas State University

Recently developed measures of clustering in free recall are examined and shown to vary with characteristics of recall unrelated to relative amounts of clustering. Specifically, previous measures were shown to vary with such factors as number of categories recalled, the distribution of the total items recalled across categories, and total number of items recalled. For these reasons, comparisons between and within subjects, as well as comparisons between experiments are most difficult. An adjusted ratio of clustering (*ARC*) measure, which does not vary with irrelevant characteristics of recall, is proposed.

The literature in recent years has shown a dramatic increase in studies of clustering in free recall stemming from recognition of the importance of these studies for understanding organizational processes in memory. Accompanying the development of this line of research has been an increasing sophistication in the measurement of clustering. The relative advantages and limitations of many of the earlier clustering indexes have been extensively discussed by Shuell (1969). This article concerns itself with an evaluation of four of the more recently developed measures and proposes another index of clustering in free recall.

Limitations of Present Measures of Clustering

Two of the measures considered here, the modified ratio of repetition (*MRR*; Bower, Lesgold, & Tieman, 1969) and the clustering (*C*) index (Dalrymple-Alford, 1970), represent modifications of the original ratio of repetition measure developed by Bousfield (1953). The other two, the Bousfield and Bousfield (1966) deviation (*BBD*) measure and the deviation (*D*) index (Dalrymple-Alford, 1970), were designed to measure amount of clustering relative to chance. These four measures suffer from two types of limitation in that (*a*) there is no fixed upper bound for *D* and *BBD*; and (*b*) the same *MRR* or *C* score may represent clustering above, at, or below chance because chance is not set at zero. More specifically, the score for perfect (i.e., maximum possible) clustering in the case of *D* and *BBD*, and the score representing chance clustering in the case of *MRR* and *C*, changes with (*a*) the number of categories that the subject recalls, and (*b*) the distribution of the total items recalled across categories (i.e., given that a subject recalls 15 items representing 3 categories, the distribution of items across categories could be 5-5-5, 3-5-7, 1-1-13, etc.). It should also be noted that the *BBD* measure is affected by a third factor, the total number of items recalled.

Because these measures do not reflect the proportion of actual category repetitions above chance to the total possible category repetitions above chance (hereafter referred to as relative amount of clustering), it is difficult to make comparisons between experiments (or between subjects). Equally important, it is difficult to evaluate the relative amount of clustering over trials for a single subject. For one set of measures (*D* and *BBD*), a positive score indicates clustering above chance, but it is impossible to determine whether the score reflects perfect or less than perfect clustering. Similarly, for the other set of measures (*C* and *MRR*), it is also difficult to ascertain the relative amount of clustering since chance may be represented by different scores. These limitations point to the need for a measure which will allow for comparison of relative amounts of clustering in different experimental situations.

The Adjusted Ratio of Clustering (ARC) Score

An index of clustering which meets these requirements has been developed in our laboratory.[3] This is an adjusted ratio of clustering (ARC) score in which chance clustering is set at zero and perfect clustering at one. In other words, the ARC score represents the proportion of actual category repetitions above chance to the total possible category repetitions above chance for any given recall protocol. Hence, the ARC score is invariant with respect to factors unrelated to relative amount of clustering. The computational formula for the ARC score is as follows:

---

[1] This research was supported in part by Public Health Service Research Grant MH-11179 to the third author.

[2] Requests for reprints should be sent to Charles P. Thompson, Department of Psychology, Kansas State University, Manhattan, Kansas 66502.

$$ARC = \frac{R - E(R)}{\max R - E(R)} \qquad [1]$$

where

R = total number of observed category repetitions (i.e., the number of times a category item follows an item from the same category),
$\max R$ = maximum possible number of category repetitions, and
$E(R)$ = expected (chance) number of category repetitions.

It should be noted that

$$\max R = N - k \qquad [2]$$

where N = total number of items recalled, and k = number of categories represented in the recall protocol. And (from Bousfield & Bousfield, 1966)

$$E(R) = \frac{\sum_i n_i^2}{N} - 1 \qquad [3]$$

where n_i = number of items recalled from Category i, and N is as before.

Comparison of Clustering Measures

This section uses illustrative recall protocols to demonstrate the stability of the ARC score relative to the other measures noted above. For purposes of comparison, the formulas for the other four measures are given below. The formulas have been written in a single notational format to facilitate clarity of exposition.

$$BBD = R - E(R) \qquad [4]$$

TABLE 1

SCORES REPRESENTING CLUSTERING MIDWAY BETWEEN CHANCE AND PERFECT CLUSTERING FOR EACH OF FIVE MEASURES FOR EACH OF EIGHT SELECTED RECALL PROTOCOLS

Protocol	*ARC*	*MRR*	*C*	*D*	*BBD*
3SE	.50	.67	.67	.33	5.00
3LE	.50	.67	.67	.33	17.00
3SU[a]	.49	.73	.67	.32	3.89
3LU	.50	.71	.67	.33	15.00
6SE	.50	.58	.58	.42	5.00
6LE	.50	.58	.58	.42	20.00
6SU[a]	.49	.67	.67	.32	3.89
6LU	.50	.65	.65	.35	17.00

[a] It was not possible to adjust the amount of clustering to a point exactly midway between chance and perfect clustering for these series.

$$MRR = \frac{R}{\max R} \qquad [5]$$

$$C = \frac{R - \min R}{\max R - \min R} \qquad [6]$$

$$D = \frac{R - E(R)}{\max R - \min R} \qquad [7]$$

where R, $\max R$, and $E(R)$ are as before, and $\min R$ = minimum possible number of category repetitions.

It should be noted that (from Dalrymple-Alford, 1970)

$$\min R = \begin{cases} 0 & \text{if } N + 1 \geq 2m \\ 2m\text{-}N\text{-}1 & \text{otherwise} \end{cases} \qquad [8]$$

where m = number of items of the type most numerous, and N is as before.

A series of hypothetical recall protocols representing clustering performance midway between chance (as defined by Equation 3) and perfect clustering (i.e., $\max R$) are given below. In these protocols, individual letters (i.e., a, b, c, etc.) represent individual items (e.g., words), and repetitions of the same letter represent repetition of different items from the same category. The eight protocols systematically vary with respect to number of categories that the subject recalls (3 or 6), total number of items recalled (18 or 54), and distribution of the total items recalled across categories (equal or unequal). Thus, the sequence designated 3SE below is a 3-

---

[3] Following preparation of this manuscript, it was learned that Gerjouy and Spitz (1966) had previously formulated the same measure of clustering. However, these authors did not discuss in detail the rationale for the development of the measure nor comment on its merits relative to other measures of clustering.

category, short list (S; i.e., 18 items recalled) with an equal number of items recalled (E) in each category. Similarly, the sequence designated 6LU is a 6-category, long list (L; i.e., 54 items recalled) with an unequal number of items recalled (U) in each category. For convenience in computation, max$R$, min$R$, $E(R)$, and R values are given for each sequence.

3SE *aaaaabacbbbbcbcccc*
$R = 10$ $E(R) = 5$ $\max R = 15$ $\min R = 0$
3LE *aaaaabbbbccccabaaaacccccabcbbbbbccccccaaabbbbccaabbaba*
$R = 34$ $E(R) = 17$ $\max R = 51$ $\min R = 0$
3SU *aacccccacbbcccccbb*
$R = 11$ $E(R) = 7.1$ $\max R = 15$ $\min R = 3$
3LU *aaaaacbcccccccabbbccccbcccccbbbbbcccaabcbcccccccaaaacc*
$R = 36$ $E(R) = 21$ $\max R = 51$ $\min R = 5$
6SE *aaaefbbbefcccedfdd*
$R = 7$ $E(R) = 2$ $\max R = 12$ $\min R = 0$
6LE *aaaaadcdbbbcbaaaaeeedbecccffeccccccddddedcbffffffbbdeb*
$R = 28$ $E(R) = 8$ $\max R = 48$ $\min R = 0$
6SU *abeeeeeedffffffedf*
$R = 8$ $E(R) = 4.1$ $\max R = 12$ $\min R = 0$
6LU *abbddddedfcccdeeeeedefdfffffcfffceffffdceeeeeffffffd*
$R = 31$ $E(R) = 14$ $\max R = 48$ $\min R = 0$

The clustering scores derived from each of the five measures for each of the eight recall protocols are given in Table 1. As can be seen, all measures except the *ARC* measure vary with characteristics of recall protocols unrelated to the relative amount of clustering. On the other hand, the *ARC* measure is invariant and indicates clustering midway between chance and perfect performance. As noted previously, the variation in the other scores results either from (*a*) an unfixed upper bound (perfect clustering score) or (*b*) failure to adjust for chance. These sources of variation may be seen clearly in Table 2 which gives, for each measure, the clustering scores that would result if the sequencing of items within each of the eight recall protocols were rearranged so that R equaled $E(R)$ (chance) or max$R$ (perfect clustering), respectively. As the table shows, the *ARC* measure is the only measure which both adjusts for chance and has a fixed upper bound.

At this point, we would like to note that Frankel and Cole (1971) were kind enough to make available a prepublication copy of their companion article appearing in this issue. In that article, they point out what they believe are weaknesses of the present (*ARC*) measure. They argue that a negative score does not mean the same thing as a positive score of the same magnitude. Apparently, they mean that a negative score of, for example, −.50 does not mean that clustering is midway between chance and *minimum* possible clustering (min$R$), whereas a positive score of .50 does mean that clustering is midway between chance and *maximum* possible clustering (max$R$). This is true.[4] However, the alternative would be to give unequal weighting to repetitions above and below chance. We believe that the *ARC* measure is more appropriate because it weights *equally* repetitions above and below chance. For example, using the 3SE protocol, it can be shown that five repetitions above chance yield a score of .50, and five repetitions below chance yield a score of −.50. For this example, each repetition above and below chance receives a score of ±.10.

---

[4] The reader will note that the maximum negative *ARC* score for any given protocol is:

$$\frac{\min R - E(R)}{\max R - E(R)}.$$

Hence, whenever the opportunity for number of repetitions below chance is not equal to the opportunity for number of repetitions above chance, the maximum negative *ARC* score will not equal −1.

We might also note that in those instances in which attention is directed toward manipulations producing clustering below chance, a slight modification of the *ARC* measure can be made to yield negative *ARC* (*NARC*) scores having the same properties as positive scores, that is, chance is fixed at 0 and the maximum lower bound is −1. The *NARC* measure is computed as follows:

$$NARC = \frac{R - E(R)}{E(R) - \min R}.$$

TABLE 2

Scores Representing Chance (Ch) and Perfect (P) Clustering for Each of Five Measures for Each of Eight Selected Recall Protocols

Protocol	*ARC*		*MRR*		*C*		*D*		*BBD*	
	Ch	P	Ch	P	Ch	P	Ch	P	Ch	P
3SE	.00	1.00	.33	1.00	.33	1.00	.00	.67	.00	10.00
3LE	.00	1.00	.33	1.00	.33	1.00	.00	.67	.00	34.00
3SU[a]	−.01	1.00	.47	1.00	.34	1.00	−.01	.66	−.10	7.89
3LU	.00	1.00	.41	1.00	.35	1.00	.00	.65	.00	30.00
6SE	.00	1.00	.17	1.00	.17	1.00	.00	.83	.00	10.00
6LE	.00	1.00	.17	1.00	.17	1.00	.00	.83	.00	40.00
6SU[a]	−.01	1.00	.34	1.00	.34	1.00	−.01	.66	−.10	7.89
6LU	.00	1.00	.29	1.00	.29	1.00	.00	.71	.00	34.00

[a] It was not possible to adjust the amount of clustering to exactly chance for these series.

Frankel and Cole also note that *ARC* scores may be skewed in their distribution, thereby violating the normality assumption of analysis of variance. However, it has been noted elsewhere that the analysis of variance is relatively insensitive to departures from normality (e.g., Winer, 1962). Furthermore, there are a number of transformations which may be used on the data to obtain normal distributions if desired (Burros, 1951).

Frankel and Cole also propose a new measure of clustering, the *Z* score. While this is one of the most mathematically sophisticated measures advanced recently, it nonetheless suffers from the same limitations of some of its predecessors. Specifically, it has no fixed upper bound. For example, using our protocols representing perfect clustering, the obtained *Z* scores were 5.31, 10.00, 4.93, 9.65, 7.52, 15.33, 5.03, and 11.56 for protocols 3SE, 3LE, 3SU, 3LU, 6SE, 6LE, 6SU, and 6LU, respectively.

In conclusion, it has been shown that the *ARC* measure is free of the limitations of previous indexes and provides an uncontaminated measure of relative amount of clustering in free recall, thereby allowing for comparisons between and within subjects and/or experiments. However, it should be emphasized that the *ARC* score is an empirical index of clustering and in no way presumes to reflect the mechanisms underlying organization in free recall. Until measures are developed which take such mechanisms into account, we recommend the use of the *ARC* score.

REFERENCES

Bousfield, W. A. The occurrence of clustering in the recall of randomly arranged associates. *Journal of General Psychology*, 1953, **49**, 229–240.

Bousfield, A. K., & Bousfield, W. A. Measurement of clustering and sequential constancies in repeated free recall. *Psychological Reports*, 1966, **19**, 935–942.

Bower, G. H., Lesgold, A. M., & Tieman, D. Grouping operations in free recall. *Journal of Verbal Learning and Verbal Behavior*, 1969, **8**, 481–493.

Burros, R. H. Three rational methods for the reduction of skewness. *Psychological Bulletin*, 1951, **48**, 505–511.

Dalrymple-Alford, E. C. The measurement of clustering in free recall. *Psychological Bulletin*, 1970, **1**, 32–34.

Frankel, F., & Cole, M. Measures of category clustering in free recall. *Psychological Bulletin*, 1971, **76**, 39–44.

Gerjouy, I. R., & Spitz, H. H. Associative clustering in free recall: Intellectual and developmental variables. *American Journal of Mental Deficiency*, 1966, **70**, 918–927.

Shuell, T. J. Clustering and organization in free recall. *Psychological Bulletin*, 1969, **72**, 353–374.

Winer, B. J., *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

(Received March 16, 1970)

Psychological Bulletin
1971, Vol. 76, No. 1, 49–57

EMPIRICAL COMPARISON OF UNIVARIATE AND MULTIVARIATE ANALYSIS OF VARIANCE PROCEDURES

THOMAS J. HUMMEL[1] AND JOSEPH R. SLIGO
University of Minnesota *Ohio University*

In this Monte Carlo study, univariate and multivariate analysis of variance procedures are used either singly or in combination to form three approaches to the analysis of multivariate normal data. Comparisons among the approaches are made with respect to the error rates per comparison, experimentwise error rates, and average number of errors in experiments with at least one error which were obtained when data generated under 36 combinations of sample size, number of variables, and proportion of variance in common were analyzed. The approach consisting of a multivariate analysis of variance followed by univariate analyses of variance is recommended as being the most useful of the three approaches.

A popular experimental design in educational and psychological research is the completely randomized design in which experimental subjects are randomly assigned to treatment groups. Following the manipulation of some independent variable, measurements are taken on a single criterion, or dependent, variable. The resulting observations are frequently analyzed using a one-way analysis of variance. In a special case of this situation, where there are two treatments, the *t* test for two independent samples is often employed.

As a variation of the above procedures it is common to use several criterion variables. For example, after assigning some subjects to a group that receives reading Method I and others to a group which receives reading Method II, an investigator may desire to gather measurements such as reading speed, comprehension, and retention.

There are at least three approaches which can be employed to analyze data from a multivariate experiment. One approach is to test each $H_0:\mu_{j1} = \mu_{j2} (j = 1 \text{ to } p)$ with a univariate analysis of variance. (In this study it is assumed that there are two groups measured on p variables.) This approach is common in the educational and psychological research literature.

A second approach to the analysis of multivariate data has been suggested by Cramer and Bock (1966). They have recommended that an overall, or multivariate, test be carried out on all the variables simultaneously by testing the hypothesis of equal population mean vectors, $H_0:\mu_1 = \mu_2$. (To this end, a statistic such as Hotelling's T^2 could be used.) This hypothesis can be written as

$$H_0: \begin{bmatrix} \mu_{11} \\ \mu_{21} \\ \vdots \\ \mu_{p1} \end{bmatrix} = \begin{bmatrix} \mu_{12} \\ \mu_{22} \\ \vdots \\ \mu_{p2} \end{bmatrix}.$$

Rejection of $H_0:\mu_1 = \mu_2$ allows one to infer that $\mu_{j1} \neq \mu_{j2}$ is the case for at least one value of j. Following a rejection, Cramer and Bock (1966) recommend that univariate analyses of variance be run on each variable separately.

Another approach has been suggested by Morrison (1967). Like Cramer and Bock (1966), he recommended that an overall test of $H_0:\mu_1 = \mu_2$ be carried out, but rather than employing univariate analyses following a rejection, he suggested using a simultaneous confidence interval procedure which was developed by Roy and Bose (1953).

The first two approaches have both been questioned (Bock & Haggard, 1968; and Morrison, 1967, respectively) as being inadequate because they do not properly account for the correlations that might exist among the p variables.

---

[1] This article is based on the first author's doctoral dissertation, which was completed August 1969 at Ohio University, Athens, Ohio.

Requests for reprints should be sent to Thomas J. Hummel, Student Personnel Office, University of Minnesota, Minneapolis, Minnesota 55455.

What effect does the correlation among the variables have on tests of hypotheses? The answer to this question involves error rates. Since there have been a number of different error rates defined by various authors, definitions of the two error rates which were of interest in this study are given here.

Following Ryan (1959), the *error rate per comparison* is defined as the probability that any given comparison will be declared significant when in fact the null hypothesis is true for that comparison. The second error rate of interest, *experimentwise error rate*, is defined for experiments containing more than one comparison and is the probability that at least one comparison will be declared significant when in fact the null hypothesis is true for all comparisons. For these two definitions it is assumed that α is constant for all comparisons.

To better understand the differences between these error rates, one can think of an experiment being replicated a large number of times. For example, an experiment is carried out on two samples, and each subject is measured on six variables. There are six comparisons. If that experiment were repeated 1,000 times, there would be a total of 6,000 comparisons. If the experimenter knew that the samples were being drawn from the same multivariate population, he would know that each time $H_0:\mu_{j1} = \mu_{j2}$ was rejected, an error had been made. If 600 rejections resulted, the error rate per comparison would be $600/6000 = .10$.

If in 100 of these 1,000 experiments three errors were made, and in another 100 experiments two errors occurred, and in an additional 100 a single error was made, errors would have been made in 300 of 1,000 experiments, and the experimentwise error rate would be $300/1000 = .30$. If all 600 errors had occurred in 100 experiments, the experimentwise error rate would be .10. (Error rates obtained in the manner described in this and the preceding paragraph would represent empirical realizations of the two probabilities defined above.)

Returning now to the question of the effect that correlation among the p variables has on testing hypotheses concerning group differences and recalling that a major function of inferential statistics is to estimate the probability of error associated with decisions, the previous question can be restated as the following: What effect does the dependency among the variables have on error rates? If the variables are correlated, the comparisons based on these variables will not be independent. When comparisons are independent it is a rather easy task to obtain the probability of encountering an error in a comparison or an experiment. The former probability equals α, and the latter is obtainable from $1 - (1 - \alpha)^p$. Whether or not the comparisons are independent does not affect the error rate per comparison (Ryan, 1959), but dependency may affect the experimentwise error rate. When univariate analyses are the only technique used to analyze multivariate experiments, the experimentwise error rate is affected and would seldom equal $1 - (1 - \alpha)^p$. The actual probability would usually be unknown (Bock & Haggard, 1968).

A researcher wishing to choose one of three approaches for analyzing multivariate data might desire answers to the following questions. How do these procedures compare with respect to the magnitude of the two error rates associated with each? In what manner are the error rates affected by the degree of correlation among the variables, number of variables involved, and sample size? The purpose of the present study is to provide at least partial answers to these questions.

Symbols employed in the Procedures section are here defined:

a = a vector of coefficients of linear combination (here a has only one nonzero element, which is equal to one);
C = lower triangular transformation matrix;
E = sample within-groups sums of squares and cross-products matrix (with elements $e_{jj'}$);
S = sample within-groups variance-covariance matrix (with elements $s_{jj'}$);
Σ = population variance-covariance matrix (with elements $\sigma_{jj'}$);
$\bar{x}_i$ = sample mean vector for the ith sample ($i = 1, 2$), with elements $\bar{x}_{ji}$ where $j = 1$ to p;
X = generated observations matrix, with elements x_{jik} and rows x'_{ik} where $i = 1, 2$; $k = 1$ to 50; and $j = 1$ to 9;
z = vector of independent, normal deviates, z_i.

PROCEDURES

The three approaches to the analysis of multivariate data which were studied empirically are presented in the following paragraphs:

Approach 1. Univariate analysis of variance. For this approach, a univariate F test was employed to test separately each of the $H_0{:}\mu_{j1} = \mu_{j2}$, $j = 1$ to p, at $\alpha = .05$. The null hypothesis was rejected at the .05 level of significance if

$$F = \frac{(\bar{x}_{j1} - \bar{x}_{j2})^2}{\dfrac{e_{jj}}{n_1 + n_2 - 2} \cdot \dfrac{n_1 + n_2}{n_1 n_2}} \geq F_{.05;\,1,\,n_1+n_2-2},$$

where $F_{.05;\,1,\,n_1+n_2-2}$ is the upper .05 point on a central F distribution for 1 and $n_1 + n_2 - 2$ degrees of freedom.

Approach 2. Combination of univariate and multivariate analysis of variance. In this second approach, a test of $H_0{:}\mu_1 = \mu_2$ was carried out using Hotelling's T^2 test where

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2).$$

$H_0{:}\mu_1 = \mu_2$ was rejected if

$$T^2 \geq \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{.05;\,p,\,n_1+n_2-p-1}$$

where $F_{.05;\,p,\,n_1+n_2-p-1}$ is the upper .05 point on a central F distribution for p and $n_1 + n_2 - p - 1$ degrees of freedom.

In addition, univariate F tests, as described above, were carried out on each variable separately. However, a rejection of $H_0{:}\mu_{j1} = \mu_{j2}$, $j = 1$ to p, at $\alpha = .05$ was allowed only if the T^2 value was significant.

Approach 3. Multivariate analysis of variance. In the third approach, a test of the equality of the mean vectors was first carried out using the T^2 test of Approach 2 above. If the T^2 value was significant ($\alpha = .05$), statistical tests were then carried out on each variable separately. For this approach, however, a simultaneous F test was employed instead of the univariate F test. Thus if

$$F_{\text{sim}} = \frac{(\bar{x}_{j1} - \bar{x}_{j2})^2}{\dfrac{e_{jj}}{n_1 + n_2 - p - 1} \cdot \dfrac{n_1 + n_2}{n_1 n_2} p} \geq F_{.05;\,p,\,n_1+n_2-p-1},$$

the $H_0{:}\mu_{j1} = \mu_{j2}$ was rejected. This decision rule, which can be used *only* in the case of two groups and where comparisons are to be made on one variable at a time, is equivalent to rejecting $H_0{:}\mu_{j1} = \mu_{j2}$ if

$$a'(\bar{x}_1 - \bar{x}_2) - \left(a'Sa\,\frac{n_1 + n_2}{n_1 n_2}\right)^{\frac{1}{2}} T_{.05;\,p,\,n_1+n_2-p-1} \leq a'(\mu_1 - \mu_2) \leq a'(\bar{x}_1 - \bar{x}_2) + \left(a'Sa\,\frac{n_1 + n_2}{n_1 n_2}\right)^{\frac{1}{2}} T_{.05;\,p,\,n_1+n_2-p-1}$$

does not cover zero. This interval provides a simultaneous .95 confidence interval for any arbitrary linear combination of p variables (Roy & Bose, 1953). For the variable-at-a-time case, the approach utilizing the simultaneous F test is equivalent to the one suggested by Morrison (1967, p. 126).

The following paragraphs contain a step-by-step description of how the study was carried out. These steps actually provide a general description of the computer program which was written for this study. For the reader who is interested, the program can be obtained from the authors.

1. A transformation matrix C was obtained which, when postmultiplied by z, transformed z into x, a random vector observation from a p-dimensional, multivariate $N(0,\Sigma)$ population, where Σ was specified by the investigator. Equations for calculating C have been presented by Moonan (1957) and Scheuer and Stoller (1962). By adding the population mean vector μ to each of the transformed vectors Cz one obtains $x = Cz + \mu$ where x is now $N(\mu,\Sigma)$. In this study $\mu_1 = \mu_2 = 0$, therefore $x = Cz$.

2. Fifty, nine-element xs were generated, and each x' became a row in X. Executing this step twice produced Sample 1 and Sample 2.

3. The two Xs (each 50 by 9) were stored. The elements of Σ were such that for both Xs, all σ_{jj} were equal and all $\sigma_{jj'}$, $j \neq j'$, were equal. Data for the following nine experiments were obtained from a pair of X matrices:

a. $n = 10$, $p = 3$
b. $n = 10$, $p = 6$
c. $n = 10$, $p = 9$
d. $n = 30$, $p = 3$
e. $n = 30$, $p = 6$

f. $n = 30, \quad p = 9$
g. $n = 50, \quad p = 3$
h. $n = 50, \quad p = 6$
i. $n = 50, \quad p = 9.$

The first three steps simulate the following situation. An investigator has two random samples of $n_i = 10$ subjects, each of which is measured on three variables. He carries out an experiment and then measures his subjects on three additional variables and carries out another experiment with $p = 6$. He repeats the process again, but this third experiment now has nine variables. Having finished three experiments, 20 additional subjects are randomly obtained for each group so that $n_i = 30$. In a manner identical to the case where $n_i = 10$, three more experiments are carried out where alternatively $p = 3$, 6, and 9. Six experiments have now been completed. Sample size is increased to $n_i = 50$. Again, three more experiments are performed where $p = 3$, 6, and 9.

Nine experiments were performed, each under a different combination of n_i and p. Each experiment was analyzed by the three approaches described previously, and the errors made by each procedure were recorded. Since some of the data in the X matrices were used in more than one experiment, the complete independence of these experiments cannot be assumed. This overlap was necessary due to cost considerations.

Only rejections of $H_0: \mu_{j1} = \mu_{j2}$ were counted. The overall test of mean vectors did not enter into the numerical calculation of error rates; rejection of $H_0: \mu_1 = \mu_2$, however, was a prerequisite for counting rejections of $H_0: \mu_{j1} = \mu_{j2}$ in the second and third approaches. This was done so that an equal number of decisions of the same type would be made by each approach. (Additionally, it is frequently of little theoretical or practical value to know *only* that two mean vectors are unequal. Often one will want to know on which variable[s] differences exist.)

4. Steps 2 and 3 were repeated 1,000 times.

5. Four sets of experiments as described in Steps 1–4 were analyzed. For each set, a different random number was used to initialize the random number generator,[2] and most important, each run was made with a different population covariance matrix. Σ was varied so that $\sigma_{jj'}^2/\sigma_{jj}\sigma_{j'j'} = \rho_{jj'}^2$, the proportion of variance in common between the jth and j'th variable, was equal to .10 in the first set, .30 in the second, .50 in the third, and .70 in the fourth.

In summary, there were 36 different combinations of levels of n_i, p, and proportion of variance in common $(3 \cdot 3 \cdot 4 = 36)$ used in this investigation, and 1,000 experiments were carried out under each combination. The error rates presented in this study are based on the number of errors which were made when each of the 36,000 experiments was analyzed by the three different approaches. (These error rates are, of course, point estimates of theoretical probabilities; however, in the following sections modifiers such as estimated and obtained are not used.)

FINDINGS AND INTERPRETATIONS

Table 1 contains the error rates per comparison and the experimentwise error rates which were obtained by analyzing samples of multivariate data with univariate analyses of variance (Approach 1). The error rates per comparison range from .040 to .055, and no consistent trend seems to exist here. Theoretically, one would expect these values to be close to .050, and this seems to be true.

In the lower half of Table 1 are the experimentwise error rates. Theoretically, in the case of $\rho_{jj'}^2 = 0$, that is, no variance in common, one should observe rates of $1 - (1 - \alpha)^p$, and for the opposite extreme of complete dependence, $\rho_{jj'}^2 = 1.0$, the theoretical value would be α. In this study, four points have been estimated between .00 and 1.00 proportion of variance in common. For purposes of comparison, the experimentwise error rates of Table 1 have been graphed in Figures 1, 2, and 3. It can easily be seen in Figure 1 that the experimentwise error rates increase as p

---

[2] The multivariate data generator used ANORM, an entry point in RANDNR, to generate the elements of z. This subroutine is in the Ohio University Computer Center library. It was programmed by Larry Irwin, Department of Mathematics, Ohio University.

increases and decrease as proportion of variance in common ($\rho_{jj'}^2$) increases. In Figures 2 and 3, based on $n_i = 30$ and $n_j = 50$, respectively, the same trends appear. Regardless of sample size, there is always less fluctuation for $p = 3$. The largest fluctuation due to variance in common occurs with $p = 9$ and $n_i = 10$. Under that circumstance the rate drops from .348 to .129 as proportion of variance in common increases from .10 to .70.

Table 2 contains the error rates obtained when a combination of multivariate and univariate techniques (Approach 2) was employed to analyze the data. The error rates per comparison are fairly homogeneous and range from .005 to .022. These values are always lower than their counterparts in Table 1. There is a consistent but very slight trend for these rates to decrease as p increases. Fluctuations are generally less than .010 and, therefore, would seem to be of little practical importance. Likewise, the experimentwise error rates are reasonably homogeneous. No consistent trends appear in the data. This is in sharp contrast to Table 1. The values in Table 2 range from .017 to .050. As in the case of the error rates per comparison, these values are always lower than their counterparts of Table 1.

TABLE 1

ERROR RATES PER COMPARISON AND EXPERIMENTWISE ERROR RATES FOR ANALYZING MULTIVARIATE DATA WITH UNIVARIATE ANALYSES OF VARIANCE

Sample size	Number of variables	Proportion of variance in common			
		.10	.30	.50	.70
		Error rate per comparison			
10	3	.054	.043	.053	.040
10	6	.054	.044	.055	.040
10	9	.053	.046	.055	.043
30	3	.044	.051	.050	.043
30	6	.047	.052	.047	.040
30	9	.046	.053	.046	.040
50	3	.052	.050	.043	.043
50	6	.049	.048	.045	.044
50	9	.051	.049	.047	.047
		Experimentwise			
10	3	.145	.112	.114	.077
10	6	.267	.190	.178	.111
10	9	.348	.247	.209	.129
30	3	.115	.119	.117	.085
30	6	.225	.200	.176	.115
30	9	.296	.263	.223	.140
50	3	.138	.124	.102	.083
50	6	.230	.190	.160	.115
50	9	.324	.258	.208	.146

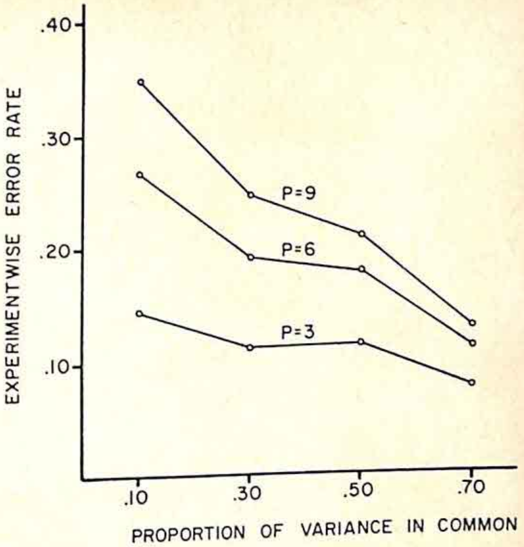


FIG. 1. Relationship between experimentwise error rate and proportion of variance in common for univariate analyses of variance ($n_i = 10$).

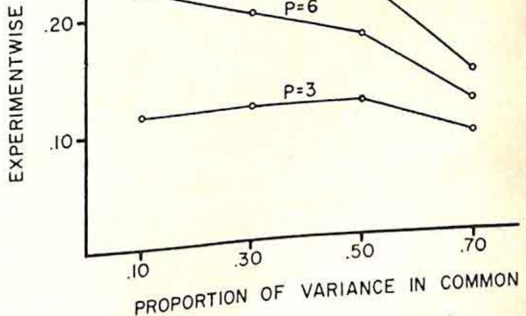


FIG. 2. Relationship between experimentwise error rate and proportion of variance in common for univariate analyses of variance ($n_i = 30$).

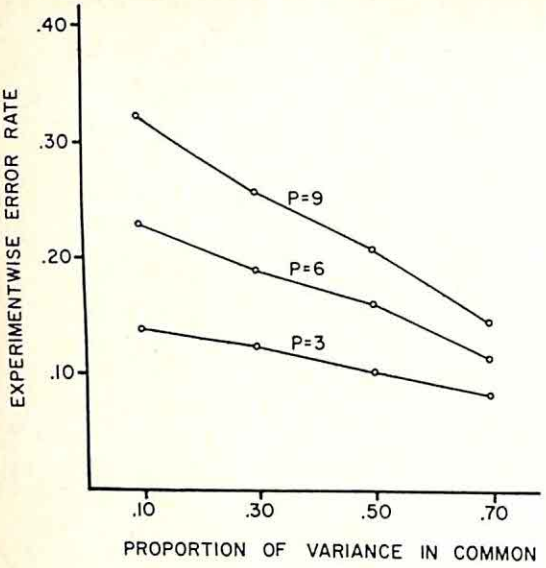


FIG. 3. Relationship between experimentwise error rate and proportion of variance in common for univariate analyses of variance ($n_i = 50$).

Table 3, for the most part, is homogenous with respect to error rates. The completely multivariate approach (Approach 3) allows for few errors. A slight but fairly consistent trend is for this approach to become more stringent as p increases. For the case of $p = 9$, 12,000 experiments were carried out. Only 1 of these experiments had an error in it. There are 72 error rates presented in Table 3. In no instances did these error rates exceed the corresponding rates in Table 2.

Table 4 was not a part of the computer output, and originally no plans were made to collect such data. However, inspection of the data suggested that it would be interesting to calculate the average number of errors made in experiments with at least one error. Only the univariate approach (Approach 1) and combination approach (Approach 2) are presented; the strictly multivariate approach (Approach 3) allows for so few errors that it was not included. With few exceptions, the average number of errors in experiments with at least one error increases as a function of both p and proportion of variance in common. For the univariate case, the highest average (2.984) was obtained under $n_i = 10$, $p = 9$, and proportion of variance equal to .70. In the combination approach, under $n_i = 50$, $p = 9$, and proportion of variance in common equal to .70, the highest average number of errors was reached, 3.950. The combination multivariate and univariate approach consistently produced higher averages than the univariate approach. In general, the discrepancy increased as p increased.

Table 5 presents the error rates for the overall T^2 test. While these rates by themselves were not of primary interest in the study, they were computed and are here presented as further evidence of the validity of Hotelling's T^2 test.

DISCUSSION AND RECOMMENDATIONS

The third approach, overall multivariate analysis of variance followed by simultaneous F tests, is discussed first and attention is then turned to the other approaches. As the number of variables increases, this completely multivariate approach becomes very conservative and would seem to be useful only

TABLE 2

ERROR RATES PER COMPARISON AND EXPERIMENTWISE ERROR RATES FOR ANALYZING MULTIVARIATE DATA WITH MULTIVARIATE ANALYSES OF VARIANCE FOLLOWED BY UNIVARIATE ANALYSES OF VARIANCE

Sample size	Number of variables	Proportion of variance in common			
		.10	.30	.50	.70
		Error rate per comparison			
10	3	.018	.012	.018	.012
10	6	.010	.009	.012	.007
10	9	.009	.007	.008	.005
30	3	.017	.022	.014	.014
30	6	.011	.013	.011	.009
30	9	.010	.012	.009	.009
50	3	.018	.021	.016	.017
50	6	.011	.013	.010	.011
50	9	.008	.011	.008	.009
		Experimentwise			
10	3	.044	.029	.035	.022
10	6	.046	.029	.030	.017
10	9	.050	.026	.025	.018
30	3	.037	.041	.029	.025
30	6	.037	.037	.032	.021
30	9	.042	.042	.030	.021
50	3	.038	.041	.033	.028
50	6	.037	.039	.028	.027
50	9	.036	.038	.026	.020

when the cost of a Type I error is extremely high. The ability of this approach to find differences when they actually exist would be questionable. If used in practice, the following anomalous situation would frequently arise. The researcher would reject $H_0: \mu_1 = \mu_2$ and infer that $H_0: \mu_1 \neq \mu_2$. However, the application of the simultaneous F tests would result in none of the $H_0: \mu_{j1} = \mu_{j2}$, $j = 1$ to p, being rejected. The researcher would thereby conclude that the vectors were unequal but that the groups did not differ on any of the variables. (It should be mentioned that this situation can also arise with the combination approach, but not as frequently.)

Assuming that a researcher does not wish to use an extremely conservative technique, he still has two approaches from which to choose: (a) univariate analyses of variance and (b) multivariate analysis of variance followed by univariate analyses of variance, the combination approach. It might be tempting to state simply that the combination approach is more conservative than the univariate approach and that a choice between the two should be based on the degree of conservatism desired. This avenue is not taken

TABLE 3

ERROR RATES PER COMPARISON AND EXPERIMENTWISE ERROR RATES FOR ANALYZING MULTIVARIATE DATA WITH MULTIVARIATE ANALYSES OF VARIANCE FOLLOWED BY SIMULTANEOUS F TESTS

Sample size	Number of variables	Proportion of variance in common			
		.10	.30	.50	.70
		Error rate per comparison			
10	3	.003	.002	.005	.003
10	6	.000	.000	.000	.000
10	9	.000	.000	.000	.000
30	3	.006	.003	.002	.004
30	6	.000	.001	.000	.000
30	9	.000	.000	.000	.000
50	3	.007	.004	.004	.003
50	6	.000	.000	.000	.000
50	9	.000	.000	.000	.000
		Experimentwise			
10	3	.010	.005	.012	.006
10	6	.000	.000	.003	.001
10	9	.001	.000	.000	.000
30	3	.017	.010	.004	.010
30	6	.000	.003	.001	.001
30	9	.000	.000	.000	.000
50	3	.022	.011	.010	.008
50	6	.002	.001	.001	.002
50	9	.000	.000	.000	.000

TABLE 4

AVERAGE NUMBER OF ERRORS IN THOSE EXPERIMENTS HAVING AT LEAST ONE ERROR

Sample size	Number of variables	Proportion of variance in common			
		.10	.30	.50	.70
		Univariate analyses of variance			
10	3	1.110	1.152	1.404	1.545
10	6	1.206	1.400	1.843	2.180
10	9	1.376	1.676	2.359	2.984
30	3	1.148	1.286	1.282	1.518
30	6	1.253	1.570	1.591	2.087
30	9	1.385	1.825	1.870	2.593
50	3	1.123	1.218	1.275	1.566
50	6	1.274	1.505	1.681	2.304
50	9	1.404	1.702	2.029	2.884
		Combination of multivariate and univariate analyses of variance			
10	3	1.205	1.241	1.543	1.591
10	6	1.326	1.828	2.433	2.529
10	9	1.600	2.385	2.840	2.722
30	3	1.405	1.585	1.414	1.640
30	6	1.757	2.135	2.063	2.571
30	9	2.143	2.571	2.633	3.714
50	3	1.395	1.537	1.485	1.857
50	6	1.730	2.077	2.071	2.444
50	9	2.000	2.526	2.846	3.950

TABLE 5

ERROR RATES FOR THE MULTIVARIATE ANALYSES OF VARIANCE

Sample size	Number of variables	Proportion of variance in common			
		.10	.30	.50	.70
10	3	.044	.040	.048	.045
10	6	.065	.050	.053	.047
10	9	.069	.046	.046	.045
30	3	.044	.056	.047	.043
30	6	.046	.051	.049	.053
30	9	.048	.055	.041	.051
50	3	.047	.053	.049	.047
50	6	.043	.058	.043	.053
50	9	.040	.046	.046	.054

here for two reasons: (*a*) the combination approach is not extremely conservative; and (*b*) there is an important difference between the univariate and the combination approaches which can serve as a basis for choice.

With the univariate approach one typically does not know the experimentwise error rate, whereas with the combination approach one can effectively estimate and control its value. There are only two instances when the experimentwise error rate is known for the univariate analysis of multivariate data, and that is when the comparisons are either completely dependent or completely independent ($\rho_{jj'}^2 = 1.00$ or $\rho_{jj'}^2 = .00$, respectively). When univariate analyses are employed and the variables are uncorrelated (given that the null hypothesis is always true), the experimentwise error rate is obtainable from $1 - (1 - \alpha)^p$. When nine independent comparisons are carried out in an experiment the error rate is .37. If the experiment were replicated 1,000 times, 370 experiments would be expected to contain at least one error. In addition, if $\alpha = .05$ for each of the 9,000 comparisons, 450 of them would be expected to be significant due to chance. On the average, there would be 450/370 = 1.216 significant comparisons in those experiments with at least one error. Obtaining three or more significant comparisons due to chance alone would be a rare event (.008).

However, this is not the case when the tests are correlated to some degree. As pointed out in Table 4, errors come in sets, and the size of these sets increases as the number of variables and the amount of dependency among the variables increase. If complete dependency exists, then all variables in an experiment are either significant or not significant. In practice, given nine variables and approximately .70 variance in common, three errors would frequently be found in experiments containing at least one error. In this instance, seeing that three tests out of nine are significant would not seem to provide any special comfort. This situation is somewhat balanced by the fact that as the proportion of variance increases, and with it the average number of errors, experimentwise error rate decreases. Even so, when the proportion of variance in common is .70 and $p = 9$ (collapsing across n_i in Table 1), the experimentwise error rate was close to .14. An average of almost three errors was made in approximately 14% of the experiments. In analyzing multivariate data with the univariate approach, a rather confusing state of affairs exists. The probability of at least one error fluctuates directly with p and indirectly with the proportion of variance in common. Experimentwise error rate seems especially important in this case because the errors occur in sets. The average number of errors also fluctuates with p and proportion of variance in common. One may well wonder whether he should be impressed when 4 tests are significant out of 11, especially if his univariate F test computer program does not provide a correlation matrix. The discussions of error rates have a certain neatness whenever independence is assumed, but introducing dependence makes the situation rather chaotic.

Compare with the above a situation in which experimentwise error rates remain relatively constant despite fluctations in p and proportion of variance in common. The combination approach offers such consistency. The errors still occur in sets, and slightly larger sets than in the univariate approach at that, but the probability of a set of errors occurring is appreciably less for this approach, and most important, the probability can be accurately estimated. For the situation described above where $p = 9$ and proportion of variance in common was .70, the average number of errors in experiments with at least one error was approximately 3.5, but the experimentwise error rate was .02 rather than .14 (again collapsing across n_i).

The above discussion leads to the following recommendations:

1. The use of the univariate approach with multivariate data should be discouraged, particularly as p and proportion of variance in common increase. The grouping of errors and the generally unknown experimentwise error rates can easily allow for misinterpretations of findings.

2. The combination approach results in an experimentwise error rate which is reasonably consistent, regardless of p and proportion of variance in common. This procedure does not eliminate the occurrence of a group of errors, but it keeps the probability of a group occurring at a low level. This approach is

recommended for its consistency and because its conservatism is not extreme.

3. The completely multivariate approach is recommended only for those who desire an extremely conservative test. In the form used in this study, this approach seemed to control for grouping of errors as a function of p, but not as a function of proportion of variance in common. This last point is stated cautiously since the extreme conservativeness of this test made it difficult to observe trends.

REFERENCES

Bock, R. D., & Haggard, E. A. The use of multivariate analysis of variance in behavioral research. In D. K. Whitla (Ed.), *Handbook of measurement and assessment in behavioral sciences.* Reading, Mass.: Addison-Wesley, 1968.

Cramer, E. M., & Bock, R. D. Multivariate analysis. *Review of Educational Research,* 1966, **36**, 604–617.

Moonan, W. J. Linear transformation to a set of stochastically dependent normal variables. *Journal of the American Statistical Association,* 1957, **52**, 247–252.

Morrison, D. F. *Multivariate statistical methods.* New York: McGraw-Hill, 1967.

Roy, S. N., & Bose, R. C. Simultaneous confidence interval estimation. *Annals of Mathematical Statistics,* 1953, **24**, 513–536.

Ryan, T. A. Multiple comparisons in psychological research. *Psychological Bulletin,* 1959, **56**, 26–47.

Scheuer, E. M., & Stoller, D. S. On the generation of normal random vectors. *Technometrics,* 1962, **4**, 278–281.

(Received February 13, 1970)

Psychological Bulletin
1971, Vol. 76, No. 1, 58–72

LINGUISTIC DEFICIENCY AND THINKING: RESEARCH WITH DEAF SUBJECTS 1964–1969[1]

HANS G. FURTH[2]

Center for Research in Thinking and Language,
Catholic University of America

Recent studies with deaf subjects are reviewed, grouped into subareas of rule learning, logical symbols, Piaget-type, memory, and perception. Comparative results on hearing controls are reported with a view to isolate the potential effects of linguistic deficiency on cognitive performance. The general conclusions of a previous review are confirmed in that the thinking processes of deaf children and adolescents are found to be similar to hearing subjects. It is pointed out that with few exceptions persons profoundly deaf from birth are severely deficient in linguistic skills in spite of many years of schooling; verbal processes could therefore not account for the emergence of cognitive skills even where developmental lags were noted. Occasional failures on certain logical tasks could be parsimoniously attributed to an unfavorable environment such as is observed in culturally different groups. The reviewed data challenge existing theories that base intellectual development largely upon linguistic learning; on the other hand, they confirm Piaget's operative theory.

Since publication of the 1964 *Psychological Bulletin* article (Furth, 1964b) a number of new studies[3] with deaf subjects have been reported. It seems that psychologists are beginning to realize the opportunity offered by the presence of linguistically deficient persons to test theories about the influence of language on various cognitive activities which are here subsumed under the word "thinking." As an illustration of the increasing popularity of this approach, seven studies with deaf children were reported in a single recent volume (1966) of a leading child psychology journal, *Child Development*.

Table 1 lists 39 studies in the order in which they are summarized in this paper. The table affords an overview and is similar, though not identical, in arrangement to Table 1 of the preceding (1964) article. The studies are roughly grouped into five subsections, and within each section put in chronological order, starting with tasks of rule learning and logical symbols, followed by Piaget-type tasks and memory tasks, and ending with perceptual tasks. In response to criticisms or misunderstandings, additional clarifications concerning the notion of linguistic deficiency are presented at the end of the present review.

Rule Learning

1. Blank and Bridger (1966) tested the hypothesis that verbal skills are a prerequisite of successful learning in situations involving absence (negation) or numbers. They observed 24 deaf and 24 hearing children, ages 3 to 4, on a concept discovery and transfer task in which the relevant cue was the presence or absence of stimulation (light or vibration). In median trials to criterion no difference was found between deaf and hearing children. In a second experiment, performances of 19 hearing and 13 deaf children, ages 5 and 6, on a successive discrimination task with visual and tactual conditions were compared. The relevant cues were the number concepts "one" and "two." Again, there were no differences in performance, except that deaf children in both experiments showed greater facility in handling tactual material than did hearing children. The authors concluded that deaf children, although deficient in general language, suc-

---

[1] Writing of this paper was supported in part by Grant RD 3111-S from Social and Rehabilitation Service, United States Department of Health, Education, and Welfare.

[2] Requests for reprints should be sent to Hans G. Furth, Department of Psychology, Catholic University of America, Washington, D. C. 20017.

[3] Excluded from this, as from the former review, are studies with deaf subjects dealing mainly with their verbal abilities or studies that require linguistic competence for comprehending instructions or giving responses.

ceeded "not because they had a non-language concept, but rather because they possessed the appropriate symbol (e.g., word, gesture) to designate the concept [p. 37]."

2. A veteran of German *Denkpsychologie,* E. Weigl, investigated another type of rule learning in the absence of verbal encoding. A series of 11 tasks was administered by means

TABLE 1

A CLASSIFICATION OF RECENT STUDIES WITH DEAF SUBJECTS

Task and age	Study: author and number	Topic	Performance of deaf in comparison to hearing subjects	
			Similar	Inferior or different
Rule learning				
3–6	1. Blank & Bridger, 1966	Cross-modal transfer	X	
4–6	2. Weigl & Metze, 1968	Rule learning	X	
4–9	3. Pufall & Furth, 1966	Double alternation	X	X (5, 6)
6–9	4. Youniss & Furth, 1966b	Discontiguity	X	
7–12	5. Andre, 1969	Reversal shift	X	
10–11	6. O'Connor & Hermelin, 1965	Visual rules	X	
9–13	7. Furth, 1964a	Combinatorial sequences	X	
6–10	8. Furth & Pufall, 1966	Combinatorial sequences	X (10)	X (6)
11–19	9. Goetzinger et al., 1967	Raven's matrices	X (18)	X (13)
17	10. Odom & Blanton, 1967	Sequential rules	X	
15–19	11. Michael & Kates, 1965	Social concepts	X	
Logical symbols				
9	12. Furth, 1966a	Logical symbol use	X	
16–20	13. Furth & Youniss, 1965	Symbol (*a*) use (*b*) discovery	X (*a*)	X (*b*)
10–14	14. Youniss & Furth, 1967	Symbol use		X
Piaget-type				
5–8	15. Youniss & Furth, 1965	Transitivity	X	
5–8	16. Youniss & Furth, 1966a	Transitivity	X	
4–12	17. Piaget, 1966	Conservation tasks	X	X (2 tasks)
9–12	18. Furth, 1966a	Quantity of liquid		X
6–9	19. Youniss, 1967	Seriation		X
11–15	20. Ross, 1966	Probability	X (15)	X (11, 13)
8–12	21. Robertson & Youniss, 1969	Anticipatory images	X	
9–11	22. Youniss & Robertson, 1970	Anticipatory images	X	X (9)
13–19	23. Furth & Youniss, 1969	Formal operational tasks		X
Memory				
6, adults	24. Lantz & Lenneberg, 1966	Color recognition	X	X (6)
7–10	25. Youniss & Furth, 1966b	Recognition strategies		X
8–12	26. Withrow, 1968	Sequence (*a*) simultaneous (*b*) successive	X (*a*)	X (*b*)
9–16	27. Rozanova, 1966	Pictures	X (15, 16)	X (9, 10)
7–15	28. Ross, 1969	Symbol span	X	
14–18	29. Goetzinger & Huber, 1964	Design reproduction	X	
13–20	30. Conrad & Rush, 1965	Type of encoding		X
12–17	31. Conrad, 1970	Type of encoding		X
Perception				
4–6	32. Oléron & Gumusyan, 1964	Embedded figures		X
6–11	33. Yashkova, 1966	Figure reversal	X (9–11)	X (6–8)
11–16	34. Gozova, 1966	(*a*) recognition (*b*) drawings	X (*a*)	X (*b*)
7–12	35. Suchman, 1966	Preference (*a*) color, (*b*) form	X (*a*)	X (*b*)
8–18	36. Carrier, 1961	Color-weight association	X	
5½–10	37. Costa et al., 1964	RT (*a*) visual (*b*) bimodal	X (*a*)	X (*b*)
Adult	38. Stoyva, 1965	EMG and REM in sleep	X	
Adult	39. Furth, 1961	Size-weight illusion	X	

of gestures; the tasks required successive pairwise selection of extreme values on a given dimension (largest, smallest; heaviest, lightest; etc.), including shifting to a different dimension on identical objects and adding or subtracting of elements in a given tower to make it in one case the highest or in another the lowest. As an example of *optimal* performance in children without knowledge of a social language, Weigl and Metze (1968) reported that 16 of 34 profoundly deaf children, ages 4 to 6½, succeeded on *all* given tasks by readily transferring the principle of extreme values to new problems. The authors considered these results to demonstrate the possibility of nonverbal logical-mathematical thinking with implications reaching beyond psychology into epistemological, linguistic, and educational theory.

3. Pufall and Furth (1966) observed yet another rule behavior which is traditionally supposed to depend on verbal cues. They administered a series of double alternation tasks that tested for generalization and stability of the acquired principle. Subjects were deaf and hearing children in an age range of 4 to 9 years ($N = 12$ at each age). At age 4 and again at ages 7 and 9 there were no differences between the groups; however, hearing children were somewhat better than deaf children at age 5 on acquisition and at age 6 on transfer.

4. Two experiments were reported by Youniss and Furth (1966b, pp. 9–13) in which 8-year-old deaf children were found to perform comparably on a simple discrimination learning task with the relevant cue and the location of the response spatially separated. At age 6, however, deaf children were not able to benefit from instructions to attend as did hearing children. With a somewhat more complex cue pattern, no differences at either age level emerged. A third experiment (pp. 16–18) concerned temporal discontiguity in an acquisition and reversal task that included familiar labelable stimuli as well as nonsense figures. The major results were as follows: (*a*) Reversal performance was easier for the 9-year-old children than for the younger group, ages 6 and 7; (*b*) the condition of 5-second delay between presentation of stimulus and response had no systematic effect; (*c*) for younger children, both hearing and deaf, familiarity of stimuli made reversal more difficult; (*d*) hearing and deaf children performed alike. These three studies, taken together, suggest that the processes which bridge spatial and temporal discontiguity are effectively similar in linguistically deficient and linguistically competent children.

5. In order to test the hypothesis that verbal mediation may facilitate reversal, Andre (1969) observed 7- and 12-year-old hearing and deaf children ($N = 96$). The results showed, as expected, that given an option a greater proportion (about 75%) of older than younger children (about 54%) chose reversal, but the main hypothesis was contradicted by two facts. The linguistically deficient deaf children performed consistently like the hearing controls, and reversal performance occurred as frequently with those hearing subjects who were able to verbalize correctly as with those who verbalized incorrectly.

6. O'Connor and Hermelin (1965) devised a series of tasks that were intended to be visual-spatial analogies of operations required in linguistic competence. These tasks were ordering or seriation, rule discovery involving both sight and touch, immediate memory, and perceptual matching. Besides psychotic and imbecile experimental groups, aphasic, deaf, and normal control groups, ages 10 to 11 ($N = 12$ in each cell), were employed. These three control groups succeeded equally well on all tasks. It was concluded that capacity to manipulate signs and symbols is not dependent on speech or verbal comprehension and that consequently the logic of thinking and competence in verbal expressions develop relatively independently.

7. Two studies focused directly on differentiation of combinatorial sequences as a crucial prerequisite for successful learning; for instance, the subject had to learn four different responses to four three-unit stimuli ordered ABB, AAB, BAA, BBA. It was hypothesized that this skill implicitly underlies linguistic competence. In both studies three groups of subjects were employed: hearing, deaf, and sensory aphasic children. The first study (Furth, 1964a) had an $N = 10$ in each group, matched on IQ, mean age of 10, and perform-

ance on a nonsequential visual paired-associates task. The sequence task was made up of four nonsense forms ABCD which generated four stimulus pairs AC, BD, AD, BC. Responses consisted of pointing to one of four cards numbered 1 to 4. A maximum of 100 corrected trials was given with a criterion of success of nine correct responses within 10 consecutive trials. With two modes of presentation the results suggested a slight retardation in the aphasic but not in the deaf group, both in simultaneous and successive presentation.

8. A second study (Furth & Pufall, 1966) did not uphold the suggested inferiority of aphasic children. Children at two age levels were studied, 6 and 10, with $N = 16$ in each group. On the visual task the younger hearing-impaired (both deaf and aphasic) children performed somewhat below the controls on successive sequences, but the older performed alike; on the simultaneous condition, no differences were noted. For hearing children, auditory successive presentation was consistently easier than visual presentation. The inferior results of the hearing-impaired groups were related to general lack of linguistic experience insofar as natural speech can provide training in the ability to master sequentially presented combinations.

9. Additional data on deaf children's performance on Raven's Standard Progressive Matrices were provided by Goetzinger, Wills, and Dekker (1967). They reported a mean of 33.3 (with a significant increase to 36.0 at retesting after 3 months) at a mean age of $13\frac{1}{2}$ ($N = 96$), an apparent lag of $1\frac{1}{2}$ to 2 years compared to the published norms. When 60 deaf children, around $18\frac{1}{2}$ years of age, were tested, two interesting results emerged. First, the Raven's score was now 45.1, well within normal limits, and second, the reading level of the older groups indicated less than 6 months advance over the younger group. The authors inferred from these results that the slight linguistic improvement could hardly explain the improved Raven's score; the poor performance of the younger age group could best be attributed to lack of general experience.

10. Odom and Blanton (1967) hypothesized that deaf children may have difficulty remembering the temporal order of unfamiliar material because sequential information is not a relevant dimension of their linguistic experience. Groups of deaf children about 17 years old and two hearing control groups (one equated on age, the other of age 11 equated on reading achievement) were shown 20 pairs of letters or of nonsense forms, in successive (SUC) or simultaneous (SIM) conditions. The sequential order of the two elements within these pairs followed the rule that five elements were always in the first place, while five others were always in the second place. Subsequently a recognition task was given which included new combinations of the familiar elements. Some of the new combinations were according to the above rule, others were reversed. It was expected that deaf persons would be less prone to make recognition errors due to sequential rule learning (systematic errors) than errors of sequential reversal (reversal errors). In terms of overall results their performance fell between the two control groups: They correctly identified 14.9 pairs (versus 15.7 and 14.1 for the two hearing groups) and made 3.1 systematic errors (versus 3.4 and 3.3) and 1.7 reversal errors (versus 1.1 and 2.0). Compared to their age controls, the deaf persons made more reversal errors particularly on SUC with nonsense forms. In sum, linguistically deficient deaf persons formulated implicit rules pertaining to sequential order in a manner similar to hearing persons but not to the same extent.

11. Michael and Kates (1965), with a selected sample of 20 deaf seniors from the Clarke School and two hearing controls, equated by age and scholastic standing, found that there were no notable differences between the groups on a discovery task which required the identification of familiar social attributes. The test material consisted of 32 picture cards portraying social situations with different affect, sex, age, or clothing. The task was to discover the pictures that were correct instances of a given "focus" picture. It was concluded that the deaf subjects in this sample manifested as mature a social attitude as the controls.

Logical Symbols

12. Chapter 14 of Furth (1966a, pp. 215–222) described a teaching demonstration of logical thinking, based on symbolized concepts of classes and logical connectives matched against pictured instances. Three groups of 16-, 13-, and 9-year-old deaf children participated. After 4 days of teaching, even the youngest deaf children succeeded in some symbolic conceptual situations. A follow-up study indicated that the performance of the 10 deaf children in the youngest age group was comparable to a class of 11 hearing peers. The following are the results in terms of mean percentage of children who succeeded on all three items of a given logical operation with hearing scores preceding those for deaf children on each type of operation: negation (73, 78); conjunction (78, 82); affirmation-simplification (82, 56); disjunction-simplification (55, 67); disjunction-exclusion (36, 33; disjunction-addition (27, 22); conjunction 1-negation (73, 44); conjunction 2-negations (45, 67); negated conjunction (18, 33); negated disjunction (27, 22).

13. The above method was derived from an experimental procedure first devised by Furth and Youniss (1965) to test logical thinking by means of a series of symbol cards (containing logical symbols for affirmation, negation, conjunction, and exclusive disjunction) and a series of instance cards illustrating specific colors or forms that corresponded or did not correspond to the symbolized expressions of the symbol card. A subject had to learn the use of the symbols by repeated matching of a symbol and an instance. The authors first established that performance on the symbol use task correlated with standard measures of intelligence in a noncollege sample of young adults; then they tested deaf young persons from a state deaf school and found their performance quite similar to the average hearing group. While this finding confirmed previous studies on logical thinking in which linguistically deficient persons did as well as hearing subjects, the same deaf persons who succeeded on symbol use were found to fail on a symbol discovery task that was comparatively easy for 13- to 14-year-old hearing youngsters, as reported in greater detail by Youniss and Furth (1964).

As the acquisition task on symbol discovery did not by itself require learning of the meaning of the symbols, success on symbol discovery could only be achieved if subjects took the initiative in discovering the meaning of the logical symbols. It is in the discovery aspect of thinking that the deaf subjects apparently failed. Following the suggestion that experiential deficiency may result in lack of intelligent initiative or motivation, the authors sampled a rural population of a lower socioeconomic strata and observed that the performance of the rural sample was similar to that of the deaf sample: They were equal to controls on symbol use, but failed on symbol discovery. Such results seem to confirm the proposition that a deaf person's failure on nonverbal thinking tasks is less related to linguistic deficiency as such than to lack of a stimulating social and experiential environment.

14. Continuing in the line of the two previous investigations, Youniss and Furth (1967) trained 24 deaf, 24 hearing culturally different (American Indians), and 24 control, young adolescents in five sessions of logic learning. After training they were tested on their ability to identify newly introduced drawings as true or false instances of symbolic concept statements. On the more complex concepts, deaf and Indian children were inferior to the control middle-class children; on simpler concepts no differences emerged. The authors distinguished between a depictive and a logical level of using symbols; they inferred that the deaf and the Indian children alike manifested the debilitating influence of long-term experiential deprivation on intellectual development.

Piaget-Type Tasks

15. Youniss and Furth (1965) investigated by nonverbal means the manifestation of the logical principle of transitivity ($A > B$; $B > C$; therefore $A > C$). Employing an inclined plane, they first taught the child that a ball of Color A would push down a ball of Color B, but not vice versa, and similarly that Ball B pushed down a third Ball C. Subsequently,

a comparable proportion of deaf and hearing children of ages 8 to 9 predicted that Ball A would push down Ball C, while hearing children at ages 5 to 6 showed no preference for any particular prediction. This performance was interpreted as indicative of the spontaneous logical order which older, but not younger, children impose on a neutral event.

16. A second study (Youniss & Furth, 1966) explored the former issue further by introducing size as an added irrelevant dimension so that in the incongruent condition a smaller ball would push a bigger ball. In both cases, however, the color of the ball was the critical variable. With hearing children in kindergarten, Grade 1, and Grade 3, it was found that "transitive" predictions were made by the older children irrespective of perceptual congruency, but the youngest age group performed according to a logical pattern only when size was congruent with logical grouping. Deaf children, comparable in age to children in Grade 3, made in general more errors, but as far as logical grouping was concerned, they performed like the hearing children. Both this and the former study were interpreted as supporting Piaget's theory of development of concrete operational thinking as largely independent of linguistic competence.

17. Piaget (1966) reported the results of a pilot developmental study carried out by F. Affolter. Twelve to 26 deaf children at various age levels were employed for each task and tested with a clay ball for judgments of conservation. Similar to results with controls, a successive order of attainment occurred in deaf children regarding substance (8 years), weight (9 years), and volume (11 years). These results indicated a slight retardation compared to a hearing control group given the same procedure; however, in conservation of surface area and volume there was a greater difference ($10\frac{1}{2}$ versus $8\frac{1}{2}$ years for surface), while conservation of quantity of liquid manifested a substantial retardation of about 3 years ($11\frac{1}{2}$ years). Piaget concluded in these words:

> The deaf children manifest a retardation in operativity that is not overly notable and the differences of retardation . . . are due rather to the procedures by which the problem is communicated. While there may exist a certain global retardation probably due to linguistic deficiency, the role of language seems to consist more in stimulating general intellectual activity and in facilitating social mobility rather than in bringing about the structures of operations [pp. 63–64].

18. Similar results and conclusions were observed by Furth (1966a, pp. 121–124; 155–158) on judgment of conservation of amount of liquid with deaf, control, and culturally different children. The results indicated a considerable time lag between the hearing and the deaf, so much in fact, that performance could not be taken as a valid manifestation of intellectual level in the deaf group. These were the results: On a logical operation which Piaget and others report as mastered by the age of 7, deaf children did not generally succeed until age 16, children in a rural area until age 11, and hearing controls until age 9. Furth suggested that these discrepancies were due primarily to differences in experience and test attitudes that influence performance on Piaget's tasks.

19. A series of investigations was reported in Youniss and Furth (1966b, pp. 32–40) and in Youniss (1967) concerning spontaneous logical rules imposed by the child in handling concrete material. The tasks were seriation, correspondence, reverse correspondence, sorting and resorting, multiple seriation, and multiple classification. In a comparison of the performances of deaf children, ages 6 to 9, with hearing controls a considerable lag in spontaneous logical behavior was observed. With additional task information and demonstration, the logical performance of deaf children approximated the level of hearing children. On the basis of these and other studies, a discrepancy between intellectual capability and performance was hypothesized; in other words, for some reason deaf children did not readily utilize logical operations that were otherwise shown to be available.

20. Ross (1966) compared hearing (ages 7 to 15) and deaf children (ages 11 to 15) on a probability task in which children were required to guess the outcome of objectively varying odds. While younger deaf children were found to be slightly poorer, the older ones, though still linguistically deficient,

caught up with the hearing children. Ross suggested that "direct exposure to and commerce with probability situations affords sufficient opportunity for the attainment of probability competence even in the absence of concomitant improvement in linguistic skills [p. 925]."

21. Two studies explored the ability to visualize transformations of physical situations in deaf and hearing children. Robertson and Youniss (1969) in a developmental study utilized some of the more recent techniques to observe what Piaget calls "mental imagery." The authors examined children's knowledge of horizontality of water level and projection of shadows at two age levels (8 to 9 and 11 to 12). The tasks included anticipation, demonstration, retest, and generalization, in this order. The 16 deaf children at each age level were similar to their controls, with younger performing consistently poorer than older children. Only on anticipation of shadows did the younger deaf children show a slight retardation.

22. A second study (Youniss & Robertson, 1970) employed two imagery tasks in which each physical transformation led to a unique result that could not be subsumed under a single linear function as in the former investigation. In the perspective task the child had to imagine what a given scene would look like from eight distinct locations (from 45° to 180°). In the rotation task a square was moved around a fixed square of a different color; the child had to imagine the location of the square as it rotated from 45° to 325°. Responses were made by selection of four alternatives. The three test phases were anticipation, demonstration with correction, and retest. Two deaf samples at mean ages 8–10 and 11–11 were matched against two hearing samples ($N = 24$ in each group). Results indicated age-related improvements. The older deaf children were comparable to their hearing peers, and the younger deaf children were similar on rotation but markedly inferior on perspective, particularly on the retest. Taken as a whole, this and the above study confirm Piaget's theory that visual forms of logical thinking do not require internalized linguistic symbols but depend primarily on the development of operational thinking structures.

23. A pilot study on formal operational thinking in deaf adolescents was presented by Furth and Youniss (1969). This is a difficult question to probe decisively by nonverbal means since the presence of formal operational thinking is primarily inferred from verbally formulated reasoning rather than from success observed in external performance. Seven deaf adolescent boys, poor in written and oral English, were selected on the basis of above average intelligence in measured IQ. They were observed on a modified nonverbal version of six Piaget "tests": combination, probability, volume, space, incline, flexibility. No one succeeded or failed clearly on all tasks, but five boys succeeded on at least one. The results were discussed in relation to Piaget's theory of formal operational thinking. Some support was seen for the proposition that formal operations develop, at least in their primary formation, without the support of linguistic competence.

Memory

24. The division of the color continuum is commonly recognized to be a more or less arbitrary affair, determined largely by social transmission. It is therefore of particular interest to observe whether linguistically deficient deaf children differ from controls on a task that requires discrimination of a particular color. Lantz and Lenneberg (1966) carried out such an investigation, using deaf and hearing subjects at age 6 and at college age. The principal task was to recall among 43 comparison colors, arranged in a circular array, two particular hues, which were first presented followed by a 5-second interval. Recognition performance was better for adults than for children and for younger hearing as compared to younger deaf children. Communication accuracy was also measured for the hearing and the older deaf groups, and positive correlations were computed between recognition and communicability Both deaf groups were found to differ somewhat in their pattern of recognition errors over the color spectrum. From these data the authors concluded that color categorization develops

somewhat differently in deaf children and that in deaf adults linguistic skill facilitates recognition without basically modifying the original patterning of colors.

25. Four pilot studies to explore strategies in recognition memory were reported by Youniss and Furth (1966b, pp. 19–25). The procedure consisted of pointing out three drawings in a presentation of nine drawings arranged in a 3 × 3 matrix. The child's requirement was to point—either immediately or after a 10-second delay—to the same three drawings when the nine drawings were shown again in a rearranged order. Recognition scores of deaf children, ages 7 to 10, were somewhat below that of hearing children; moreover, with increasing age, hearing children pointed to the three drawings in correct temporal order with greater accuracy, particularly when familiar (hence labelable) objects were involved. A final experiment included a preexposure period to facilitate attention to the two-dimensional relation of the matrix—from smaller to bigger and from lighter to darker. Under these conditions the scores of all the children, but especially the deaf children, improved, and deaf-hearing differences became negligible. Deaf children demonstrated inferior recognition performance since they did not employ as effective strategies as did hearing children. The readily induced effect of training seemed to indicate that this inferior recognition was not due to a deficit in intellectual structure but to a lack of spontaneous utilization of strategies.

26. Withrow (1968) addressed himself to a question also taken up by investigators mentioned before (Studies 6, 7, 8, 10), namely, the relation of sequence learning and linguistic deficiency. He tested on recall of simultaneously (SIM) or successively (SUC) presented sequences four groups of matched children, ages 8 to 12: three groups of 14 deaf children, each taught by different methods, and 14 hearing children. Stimulus material were three classes of visual forms, in order of decreasing meaningfulness: silhouettes of familiar objects, geometric forms, and random forms. Number of items per sequence varied from two to six, exposure times for SIM were $\frac{1}{16}$ or $\frac{3}{16}$ seconds, and rates of presentation for SUC were one to four forms per second. Three main results were reported: (*a*) Number of items per trial, meaningfulness, and exposure time affected all groups in the expected direction. (*b*) The three deaf groups did not differ among themselves. (*c*) Hearing children were superior to deaf in recall of SUC, but on SIM there was no difference. For the deaf groups, mean SUC and SIM scores were similar; for the hearing groups, SUC was easier. In conclusion, some measurable difference in the deaf child's immediate memory for successive items was observed in the form of a slight retardation that could easily be attributed to lack of linguistic experience.

27. Rozanova (1966) reported two experiments on the pictorial memory of deaf children. In Experiment 1, partial and whole pictures were first presented and had to be reproduced after 10 minutes. Younger deaf children (ages 9 to 10) were found to perform more poorly than hearing controls; the deficiency was more marked in the partial condition and disappeared with age, resulting in no differences at ages 15 to 16. In a second experiment, deaf and hearing children (ages 8, 10, and 12) were shown 16 pictures of familiar objects and had to recognize them 20 minutes later within a set of 80 cards including some similar "trick" cards. Younger age and deafness were found to affect memory performance adversely. On the same task, after a further interval of 10 minutes, two or three cards at a time were shown to the child, including incorrectly chosen pictures. In this condition deaf children corrected their mistakes as well as hearing children.

28. In a test of memory for a span of nine binary symbols (+, −), Ross (1969) serially presented *patterned* series (alternating symbols with the insertion of a run of three identical symbols) and *unpatterned* series (alternations and doubles in random sequence). Patterned and unpatterned series composed of three or four different symbols were also administered. Deaf and hearing children at the five ages tested (7 to 15) did not differ in terms of errors, and both groups performed better on the same sorts of patterned series. In opposition to results with

digit span, average span length was only about four symbols for all unpatterned series. The similarity of results between deaf and hearing for all types of series challenges some recent theoretical positions that require an intact auditory-vocal system for normal retention of sequential visual information.

29. Goetzinger and Huber (1964) matched 30 hearing and 30 deaf youngsters on age (14 to 18) and tested them on memory-for-design (Benton Visual Retention Test). On immediate reproduction both groups were similar, both being somewhat worse with a 5-second than with 10-second presentation. But on 10-second presentation, followed by 15-second delay, hearing children experienced no decrement in their performance due to delay; deaf children, however, performed worse, particularly when the delay condition came first. The authors suggested that verbalization may have aided the hearing children in memorizing the design when a delay was interspersed between presentation and reproduction.

30. In order to examine order errors indicative of type of memory encoding, Conrad and Rush (1965) presented five-letter sequences from a given list of 10 consonants for immediate written recall to a sample of deaf youngsters ($N = 41$, age range 13 to 20). From a control group of 53 hearing seventh graders, a subsample of the worst 15 was matched with a subsample of the best 15 deaf subjects in terms of similar mean memory performance (71.7% retained letters). For the best as well as the worst hearing subsamples, significant rank order correlations (.47 and .52) emerged between memory errors and acoustic listening errors established previously by Conrad. Deaf subjects also made some systematic errors, but these were neither acoustically nor spatially related to correct letters. This pilot study thus confirmed the acoustic encoding of letters for hearing subjects and its absence for deaf subjects, with the mode of deaf encoding processes still unclarified.

31. In a follow-up investigation (Conrad, 1970) 36 deaf boys read, silently and aloud, five- or six-consonant sequences from a list of nine consonants with exposure time of 5 or 6 seconds, respectively. The task require immediate written recall of each sequenc On the basis of obtained confusion errors, 2 subjects with typical articulatory errors similar to hearing controls were classified int Group A, and the rest with errors perhap related to shape into Group non-A. Ratings of speech quality correlated with this classification. In contrast to hearing subjects who regularly benefit from reading out loud over silent reading, notably by a stronger recency effect, Group A was unaffected by reading modes, and the Group non-A made even more recall errors with reading out loud. A second experiment demonstrated that Group A had more difficulty in serial learning of a list of homophone word pairs than a list of pairs of words of similar shape; whereas Group non-A learned both lists equally well. This study provided interesting evidence of different coding types on the part of deaf children. But even those deaf subjects who appeared to use auditory-articulatory encoding were not as uniformly or strongly affected by auditory-related conditions as are hearing persons.

Perceptual Tasks

32. Oléron and Gumusyan (1964) employed a perceptual task of embedded figures (Poppelreuter Test) and tested 28 deaf and 28 hearing children at ages 4 to 6. The children had to choose from a series of individual drawings those objects or geometric designs which they recognized in the composite picture of superimposed drawings. With a scoring system that rewarded correct choices and penalized wrong choices, deaf children were consistently somewhat poorer; however, only the 5-year-olds (on the meaningful figures task) emerged significantly poorer in statistical comparison. As the deaf appeared perceptually inferior, not only on meaningful (hence nameable) objects but also on geometric designs (for which names were not available to the hearing children), the authors suggested that linguistic skill may facilitate the establishment of an interior symbolic habit which intervenes in the perceptual recognition of the environment.

33. Yashkova (1966) published a report on deaf children's ability to make reverse

drawings of geometric figures. Deaf children, ages 9 to 11, were found to be as good as hearing controls, while at younger ages deaf children were somewhat retarded. Additionally, the author reported overall improvement following instructions in generalized methods of figure reversal. In early trials, spontaneous discovery of a generalized method was observed in hearing children at 5 years; with deaf children, however, only at 8 years.

34. Spatial notions of deaf children, Grades 5 to 10, were investigated by Gozova (1966). In recognition tasks young children were provided with a schematic drawing of a three-dimensional form which they had to find among a number of concrete objects. Older children had to recognize an object by its drawing in three dimensions and then make a drawing of it. The observed similarity of performance on the first task suggested that spatial notions are normally developed in deaf children while the observed retardation on the second task indicated to the author a lack of attention and detailed thinking on the part of older deaf children.

35. As an extension of earlier work by Doehring (1960) on color-form attitude, Suchman (1966) replicated Doehring's test on 36 hearing and 36 deaf children with a mean age of 10. She observed that 86% of the hearing children and only 33% of the deaf children preferred form, while only 14% of hearing children and 67% of deaf children preferred color. Subsequently these same children manifested a similar difference in discriminative accuracy: hearing children had higher accuracy scores for form than deaf children, and, conversely, deaf were better than hearing children for color discrimination. On a third task, discrimination learning, hearing excelled over deaf children on the form test, while on the color learning task both groups were equal. On the basis of these data, Suchman viewed the color preference of deaf children as an acquired habit of categorization that is more adequate vis-à-vis color cues than form cues.

36. Carrier (1961) investigated a question related to the one asked by Lantz and Lenneberg (1966) as well as by Suchman (1966). He found that hearing children over a wide age range (6–18 years) showed a significant tendency to judge a black cube heavier than a white cube of similar size and material. The converse, that white is judged lighter in weight than black, was also observed, although to a lesser degree. In a comparison of other colors, yellow was reacted to like white, while blue and red elicited no systematic weight reaction. The experimental procedure was essentially nonverbal and included a training session with a balance and weights. Deaf children over age 12 performed consistently similar to controls, but younger deaf children, below age 10, failed to manifest systematic color-weight associations. Subsequently, Carrier observed that hearing children rated color words in a weight hierarchy with black as heavy and white as light. The author suggested on the basis of these results that color-weight associations are primarily mediated by general linguistic experience rather than by the arbitrary fact that in English the word "light" combines the two elements of the weight-color association in one word.

37. Costa, Rapin, and Mandel (1964) investigated reaction time to visual, auditory, and paired (visual and auditory) stimulation in a group of 30 children with communication disorders and 28 control children of mean age 8 years. It was established that visual reaction time was similar in both groups and that positive reinforcement improved performance on successive days. For controls, auditory reaction time was shorter than visual, and paired reaction time was still shorter than auditory reaction time. In deaf children, however, paired stimulation had no facilitating effect.

38. It has long been held that with deaf persons who habitually use manual means of communication one can readily observe finger electromyography (EMG) movement, analogous to something like internal verbalization. Moreover, rapid eye movements (REMs) have recently been found associated with dreaming. Consequently Stoyva (1965) investigated whether in deaf adults REM during sleep would be related to EMG to a significant degree. On the basis of two experiments during sleep, one with seven deaf adults who habit-

ually use manual finger spelling and the other with six deaf and six hearing controls, the following results were established. In deaf persons REM, but not EMG, was associated with dreaming; EMG activity during REM periods in deaf was similar to that in hearing adults, and during non-REM periods there was a high degree of overlap between deaf and hearing groups. These findings by means of modern electrical techniques seem to challenge the former view that linked finger EMG to image activity in deaf persons.

39. Furth (1961) found deaf adolescents to be as prone to the size-weight illusion as hearing peers. Moreover, deaf as well as hearing subjects could be trained to overcome in a large measure the illusion and to show the permanent effect of training on a trial a week later.

Conclusions

The original summary article (Furth, 1964b) was criticized by Blank (1965). She contended that Furth presented inadequate evidence for stating that thinking can develop without the benefit of language and pointed out that many deaf children are well trained in language and thus cannot be considered linguistically deficient. Clarifications on these and other critical points were incorporated in a book in which Furth (1966a) presented a more comprehensive treatment of his view from a theoretical angle. This view is closely patterned after Piaget's developmental theory (Furth, 1966c).

It would be appealing to the theorist if one could single out specific cognitive areas where deaf children perform similarly to hearing children and others where they differ or fail. Alternatively, one may perhaps be in a position to summarize the data by separating from other tasks those (Studies 5, 8, 20, 22, 27, 33) on which younger deaf children showed a deficit but at a later age "caught up." Unfortunately, a glance at Table 1 in this and the former review does not suggest a single, uniform answer to either alternative. The investigations of logical symbol tasks illustrate that generalizations based on isolated studies are frequently premature. Studies 12 and 13 suggested to the investigator that deaf children were poor in *discovery* of logical symbols but equal in the intelligent *use* of symbols that were shown to them. Yet Study 14 reported a significant developmental delay in symbol *use* for deaf subjects compared to middle-class hearing controls; at the same time, results for the deaf were the same as those for a group of Indian adolescents.

Moreover, a haunting question always remains in cases where deaf children eventually do catch up: Would a more stringent criterion of success or a more complex task yield differences? An analogous question can be raised with respect to all tasks on which significant failures were observed: Would some training raise the performance of deaf subjects to a level equal to that of hearing controls, and what duration and intensity of training would be required? In order to arrive at a more conclusive answer to these questions this investigator and his associates are presently engaged in a detailed study of selected deaf adolescents which includes short-term and long-term training with individual observations and uses hearing control groups from different social environments.

Before attempting to suggest some tentative theoretical conclusions, it may be appropriate to restate in condensed form the basis on which the cited empirical studies with deaf subjects are evidence for these conclusions.

The assertion that deaf persons are linguistically deficient does not mean that they are lacking in general symbolic ability or that they are ignorant of any and all English words. Most deaf persons in our society know some English words or phrases; but admittedly the most vital aspect of the living language is not single words but the structure of the language, into which single words are fitted to form meaningful sentences. If language is said to influence thinking, this should refer not so much to single words as to the syntactical and grammatical structure of the language. It is precisely this general structure that hearing children assimilate with relative ease and the vast majority of deaf children fail to attain with an adequate degree of competence.

Commonly, the fact is overlooked and comes as a surprise to many that the majority of people born deaf, even though they spend 10 to 15 years in schools with almost single-minded emphasis on language learning, do not have competence in the natural language of their society. For understandable reasons this fact is usually not broadcast, and, with the exception of one normative study which omits comparisons with hearing norms, only summarizing statements to this effect by educators of the deaf can be gleaned from the literature.

A recognized differential measure of linguistic competence does not exist. An additional difficulty with profoundly deaf persons is that a comfortable skill in speech or lip reading is an exceedingly rare achievement. As a reasonable compromise, comprehension of fourth-grade reading as measured by present standardized tests can be suggested as an appropriate criterion. Reading tests below Grade 4 sample only fragmentary aspects of language, and there should be general agreement that knowing a few words and isolated phrases does not constitute linguistic competence. Significantly, the standardizing study of Wrightstone, Aronow, and Moskowitz (1963), based on an imposing norming total of 5,824 deaf pupils between the ages $10\frac{1}{2}$ and $16\frac{1}{2}$, demonstrated strikingly that the average reading achievement of deaf pupils stops short at Grade 3. On the Metropolitan Elementary Reading Test the youngest age group (ages $10\frac{1}{2}$ to $11\frac{1}{2}$, $N = 654$) had a mean raw score of 12.6, barely above the chance level of 11, while the respective standard deviations were 8.14 and 9.51. In terms of national norms this indicates (Furth, 1966b) that the mean score over a span of 5 years increased from a grade equivalent of 2.8 to Grade 3.6, never reaching the mean raw score of 25, the norm for beginning Grade 4. Moreover, hardly 10% of the total sample scored above Grade 4, a reading level that experts consider to fall within "functional literacy." It stands to reason that the poor reading performance of deaf school children cannot be relegated to any of the usual types of reading difficulty. A slow reader is rightfully assumed to be quite competent in the language in which he has reading difficulties. Not so with deaf persons. In their case, reading level is almost uniquely determined by competence in language and therefore is diagnostic of that competence.

Here, then, is quantified evidence that only a small percentage of deaf persons are at all comparable to hearing persons in being "at home" in the conventional language of the society. In other words, apart from a few notable exceptions, the difference in linguistic skill between persons born profoundly deaf and hearing persons is not merely statistically significant, but it is as great as the difference in level of hearing. At one extreme, there are hearing youngsters steeped from earliest childhood in the conventional language; at the other extreme, there are deaf youngsters who during the early formative years of their life have no ready-made symbol system available. The conventional gestural sign language in use among deaf adults is only available in childhood to those deaf children who are born of deaf parents—a very small percentage of the deaf population. Hence, it is literally true that most deaf children grow up during the early formative period of life without any systematized conventional language. The significant fact is that in spite of not having been provided with a conventional symbol structure, deaf children construct their own symbols as they are needed for the development of thinking. On this crucial fact rests the present writer's interpretation of the reviewed studies.

Another important point for a theoretical interpretation is understanding what can and what cannot be demonstrated by means of a deficiency experiment. If an organism without a certain factor fails on a given task, one cannot conclude that the factor is directly related to the task since other uncontrolled influences may be at work. But if he succeeds on a given task, one can infer conclusively that the missing factor is not a prerequisite for the task. As a consequence, those experiments where deaf subjects are somewhat retarded produce a far less conclusive interpretation than experiments where they are equivalent.

The apparently well-founded objection that the proposed interpretation regards deaf subjects as forming a single homogeneous group in linguistic deficiency does not quite touch the core of the argument. Unquestionably, in studying deaf children, account must be taken of wide individual variations in degree and type of hearing loss and in psychological adjustment to the loss. However, the argument presently proposed is not based on a disregard of individual differences but on the factor of linguistic deficiency that is common to the vast majority of individuals within the deaf sample. Actually some of the individual differences, such as possible neurological lesions associated with hearing impairment, would be working against the intended null hypothesis.

Another objection to the present interpretation points to several instances of disappearance at a later age of hearing-deafness differences observed at a younger age level. A possible explanation is that older deaf children are linguistically more competent than younger deaf children. Such an explanation would be more plausible if the linguistic increase were truly substantial. However, it was shown that even at a near adult age most deaf persons are still woefully lacking in linguistic competence. Whereas the observed data are clearly consistent with the postulation of a certain developmental lag for deaf in comparison to hearing children, to suggest that the catching up is due to a slight increase in linguistic competence is inadequate to explain substantial increases that take place in a number of quite diverse tasks. Rather, most aspects of deaf children's time lag can be more parsimoniously attributed to an easily demonstrable restricted past experience and training, not unlike the situation that prevails for culturally different hearing children.

During the Moscow International Congress of Psychology, O'Connor (1966) and Furth (1966d) presented summaries of recent work with deaf subjects. O'Connor reported on a number of studies intended to demonstrate the importance of speech for thinking. Yet, he concluded, the recent "findings . . . force on me the somewhat unwelcome conclusion as far as I am concerned that whatever effect auditory deprivation may have on speech, i cannot be said to have a strong effect on reasoning capacity [p. 38]."

The recent work in deafness has produced further confirmation of the theoretical implications presented before. The data confirm that at least up to and including the concrete operational stage linguistic support is not required. Concerning the relation of language to formal operational thinking, data are still sketchy, and firm conclusions are hard to draw, if only because these operations are not easily tested even in the adult hearing population. However, the general conclusion seems abundantly clear that thinking processes of deaf children are similar to those of hearing children and therefore must be explained without recourse to verbal processes. This interesting fact implies that the current emphasis in psychology that attributes the growth of intelligence mainly to verbal factors is unquestionably inadequate and that undue emphasis on linguistic skill, particularly in early education, may not be a sound basis for nourishing intellectual development (Furth, 1970). Insofar as Piaget (Inhelder & Piaget, 1964, p. 293) stated categorically that language is not a constitutive element of logical thinking, it is intriguing to consider that the evidence collected in this review provides as strong empirical support as can be reasonably expected for his theory. As a general approach, nonverbal methods may be extremely fruitful for an objective study of thinking and for an adequate appraisal of the function of the conventional language and other symbols in the development of thinking.

REFERENCES

Andre, J. Reversal-shift behavior and verbalization in two age groups of hearing and deaf children. *Journal of Child Experimental Psychology*, 1969, 7, 407–418.

Blank, M. The use of the deaf in language studies: A reply to Furth. *Psychological Bulletin*, 1965, **63**, 442–444.

Blank, M., & Bridger, W. H. Conceptual cross-modal transfer in deaf and hearing children. *Child Development*, 1966, **37**, 29–38.

Carrier, E. O. *The influence of language in the color-weight associations of hearing and deaf children.* (Science Education Research Report) Cambridge: Harvard Graduate School of Education, 1961.

Conrad, R. Short-term memory processes in the deaf. *British Journal of Psychology*, 1970, **61**, 179–194.

Conrad, R., & Rush, M. L. On the nature of short-term memory encoding by the deaf. *Journal of Speech and Hearing Disorders*, 1965, **30**, 336–343.

Costa, L. D., Rapin, I., & Mandel, I. J. Two experiments in visual and auditory reaction time in children at a school for the deaf. *Perceptual and Motor Skills*, 1964, **19**, 971–981.

Doehring, D. G. Color-form attributed of deaf children. *Journal of Speech and Hearing Research*, 1960, **3**, 242–248.

Furth, H. G. Effect of training on the adaptation level of the size weight illusion with normal, deaf, and blind subjects. *Perceptual and Motor Skills*, 1961, **13**, 155–160.

Furth, H. G. Sequence learning in deaf and hearing children. *Journal of Speech and Hearing Disorders*, 1964, **29**, 171–177. (a)

Furth, H. G. Research with the deaf: Implications for language and cognition. *Psychological Bulletin*, 1964, **62**, 145–164. (b)

Furth, H. G. *Thinking without language: Psychological implications of deafness*. New York: The Free Press, 1966. (a)

Furth, H. G. A comparison of reading test norms of deaf and hearing children. *American Annals of the Deaf*, 1966, **111**, 461–462. (b)

Furth, H. G. Language et pensée opératoire: Conséquences tirées des sourds adultes. *Bulletin de Psychologie*, 1966, **19**, 673–676. (c)

Furth, H. G. Development of thinking in the deaf: Implications for the relation of thinking and language. In R. M. Boskis & A. I. Meshcheryakov (Eds.), *Mental development and sensory defects*. Symposium 33. Moscow: International Congress of Psychology, 1966. (d)

Furth, H. G. *Piaget for teachers*. Englewood Cliffs, N. J.: Prentice-Hall, 1970.

Furth, H. G., & Pufall, P. B. Visual and auditory sequence learning in hearing-impaired children. *Journal of Speech and Hearing Research*, 1966, **9**, 441–449.

Furth, H. G., & Youniss, J. The influence of language and experience on discovery and use of logical symbols. *British Journal of Psychology*, 1965, **56**, 381–390.

Furth, H. G., & Youniss, J. Thinking in deaf adolescents: Language and formal operations. *Journal of Communication Disorders*, 1969, **2**, 195–202.

Goetzinger, C. P., & Huber, T. G. A study of immediate and delayed visual retention with deaf and hearing adolescents. *American Annals of the Deaf*, 1964, **109**, 293–305.

Goetzinger, C. P., Wills, R. C., & Dekker, L. C. Non-language I.Q. tests used with deaf pupils. *Volta Review*, 1967, **69**, 500–506.

Gozova, A. P. Spatial notions of deaf schoolchildren. In R. M. Boskis & A. I. Meshcheryakov (Eds.), *Sensory defects and mental development*. Symposium 33. Moscow: International Congress of Psychology, 1966.

Inhelder, B., & Piaget, J. *The early growth of logic in the child*. New York: Harper & Row, 1964.

Lantz, D., & Lenneberg, E. H. Verbal communication and color memory in the deaf and hearing. *Child Development*, 1966, **37**, 765–780.

Michael, J., & Kates, S. L. Concept attainment on social materials by deaf and hearing adolescents. *Journal of Educational Psychology*, 1965, **56**, 81–86.

O'Connor, N. Sensory defects and mental development. In R. M. Boskis & A. I. Meshcheryakov (Eds.), *Sensory defects and mental development*. Symposium 33. Moscow: International Congress of Psychology, 1966.

O'Connor, N., & Hermelin, B. Visual analogies of verbal operation. *Language and Speech*, 1965, **8**, 197–207.

Odom, P. B., & Blanton, R. L. Rule learning in deaf and hearing subjects. *American Journal of Psychology*, 1967, **80**, 391–397.

Oléron, P., & Gumusyan, S. Analyse perceptive et language: Application d'une épreuve de Poppelreuter à des enfants sourds et entendant. *Psychologie Francaise*, 1964, **9**, 47–60.

Piaget, J. Surdi-mutité et conservations opératoires. In, *Etudes d'epistémologie génétique* (No. 20). Paris: Presses Universitaires de France, 1966.

Pufall, P., & Furth, H. G. Double alternation behavior as a function of age and language. *Child Development*, 1966, **37**, 653–662.

Robertson, A., & Youniss, J. Anticipatory visual imagery in deaf and hearing children. *Child Development*, 1969, **40**, 123–135.

Ross, B. M. Probability concepts in deaf and hearing children. *Child Development*, 1966, **37**, 917–928.

Ross, B. M. Sequential visual memory and the limited magic of the number seven. *Journal of Experimental Psychology*, 1969, **80**, 339–347.

Rozanova, T. V. Pictorial memory of deaf children. In R. M. Boskis & A. I. Meshcheryakov (Eds.), *Sensory defects and mental development*. Symposium 33. Moscow: International Congress of Psychology, 1966.

Stoyva, J. M. Finger electromyographic activity during sleep: Its relation to dreaming in deaf and normal subjects. *Journal of Abnormal Psychology*, 1965, **70**, 343–349.

Suchman, R. G. Color-form preference, discriminative accuracy and learning of deaf and hearing children. *Child Development*, 1966, **37**, 439–452.

Weigl, E., & Metze, E. Experimentelle Untersuchungen zum Problem des nicht sprachgebundenen begrifflichen Denkens. *Schweizerische Zeitschrift für Psychologie und ihre Anwendungen*, 1968, **27**, 1–17.

Withrow, F. B. Immediate memory span of deaf and normally hearing children. *Exceptional Children*, 1968, **35**, 33–41.

Wrightstone, J. W., Aronow, M. S., & Moskowitz, S. Developing reading test norms for deaf children. *American Annals of the Deaf*, 1963, **108**, 311–316.

YASHKOVA, N. V. Intellectual reversibility formation in deaf and in normal children. In R. M. Boskis & A. I. Meshcheryakov (Eds.), *Sensory defects and mental development*. Symposium 33. Moscow: International Congress of Psychology, 1966.

YOUNISS, J. Psychological evaluation of the deaf child: Observations of a researcher. *The Eye, Ear, Nose and Throat Monthly*, 1967, **46**, 458–464.

YOUNISS, J., & FURTH, H. G. Attainment and transfer of logical connectives in children. *Journal of Educational Psychology*, 1964, **55**, 357–361.

YOUNISS, J., & FURTH, H. G. The influence of transitivity on learning in hearing and deaf children. *Child Development*, 1965, **36**, 533–538.

YOUNISS, J., & FURTH, H. G. Prediction of causal events as a function of transitivity and perceptual congruency in hearing and deaf children. *Child Development*, 1966, **37**, 73–82. (a)

YOUNISS, J., & FURTH, H. G. *Spatial and temporal factors in learning with deaf children: An experimental investigation of thinking.* (Vocational Rehabilitation Administration Report RD-1305-S) Washington, D. C.: Catholic University of America, 1966.

YOUNISS, J., & FURTH, H. G. The role of language and experience on the use of logical symbols. *British Journal of Psychology*, 1967, **58**, 435–443.

YOUNISS, J., & ROBERTSON, A. Projective visual imagery as a function of age and deafness. *Child Development*, 1970, **41**, 215–234.

(Received February 17, 1970)

Vol. 76, No. 2 AUGUST 1971

Psychological Bulletin

Diary No. ... Date 4-11-71

PSYCHOPHYSIOLOGICAL APPROACHES TO THE INVESTIGATION OF DESENSITIZATION AND RELATED PROCEDURES[1]

ANDREW M. MATHEWS[2]

University of Oxford

All available studies concerning desensitization and utilizing psychophysiological measurement are reviewed and evaluated. Autonomic changes, particularly in heart rate and skin resistance, do occur following desensitization, and this treatment seems a relatively effective method of modifying such physiological components of fear. The reported effects of phobic imagery are consistently different from those of neutral imagery and show a characteristic response decrement with repetition, although the role played by relaxation in facilitating this decrement remains unclear. Partial dissociation between physiological and other behavioral changes in treatment appears frequently, and may be attributed to varying rates of change in the different systems. Some theoretical accounts of the mechanisms involved in desensitization are discussed in light of the evidence reviewed.

The importance of autonomic responses in the genesis and maintenance of fears and phobias is frequently taken for granted, despite our continuing ignorance of the precise nature of the relationship between the physiological and psychological components of fear. Current opinion based on what evidence is available appears to be converging on the view that fear, normal or abnormal, may most usefully be viewed as a label referring to the simultaneous occurrence of responses, physiological and psychological, in a series of imperfectly coupled systems, for instance, the verbal/cognitive, behavioral/motor, and physiological/autonomic systems (Lacey, 1967; Lang, 1969; Lazarus, 1968).

While the nature of the interactions involved is not entirely agreed upon, it has become forcibly apparent that assessments of fear derived from different measuring devices, although related to some degree, show a marked degree of independence from one another. As Lang has pointed out, there are no a priori reasons why we should assume that any of the different attributes of fearful behavior exclusively define the presence or absence of fear—rather we should assess as many components as possible and take them all into account. The greater the degree of relationship between them, the greater the likelihood, and usefulness, of our labeling the observed state "fear."

This view may also be applied to processes directed towards the elimination of fearful behavior with potential profit. While psychotherapists are inclined to direct their attentions exclusively to modification via verbal expressions of emotion, and may thus influence other systems indirectly (e.g., Lacey, 1959), behavior therapists are characterized by a readiness to directly attempt the induction of behavioral and autonomic changes as well.

While this is true of many procedures called behavior therapy, such as aversion, only systematic desensitization treatment is considered here, partly because of the more wide-

---

[1] The present review was completed while the author was supported in part by the Medical Research Council, United Kingdom. An extended version was submitted in partial fulfillment of the requirements for a PhD degree at the University of London.

[2] Requests for reprints should be sent to Andrew M. Mathews, Department of Psychiatry, Psychological Treatment Research Unit, Warneford Hospital, University of Oxford, Oxford, England.

© 1971 by the American Psychological Association, Inc.

spread clinical use and application of this technique, and partly because the treatment was devised and formulated with the modification of behavioral and autonomic fear responses in mind (Wolpe, 1958). Despite this fact and the considerable interest aroused by the apparent success of desensitization, until recently it was not possible to say whether or not this treatment had any specific effects on autonomic or muscular responses. Increasing experimentation in this field is beginning to throw some light on the nature of the physiological events accompanying and resulting from desensitization, and on their interaction with the psychological changes occurring simultaneously.

Psychophysiological measurement may be used to investigate the process or outcome of desensitization for a number of different purposes, varying from immediate clinical usefulness through to fundamental investigations of the processes underlying behavioral change. For the sake of convenience, the studies reviewed here are divided into four major areas, which are outlined below.

1. Physiological variables may be monitored in the course of treating a patient or client with desensitization. The use of any one physiological measure in this way may be determined by underlying and often mistaken assumptions about the relationship between activity in that measure and anxiety—typically the skin resistance response is naively used as an "index of anxiety."

The motive for the introduction of such a measure may simply be to assist treatment, or as a clinical study to investigate physiological changes in the course of such treatment. Either way, most such studies must be considered uncontrolled in that appropriate control groups are not included, and the procedures used are often unstandardized in response to the demands of treatment in the individual patient.

Nevertheless, these studies may have some distinct advantages; they remain close to the clinical treatment situation and usually concern seriously disturbed subjects, and thus can be a rich source of hypotheses which may be tested later in controlled studies.

2. In controlled studies contrasting desensitization with alternative treatment procedures, it seems appropriate to utilize a wide variety of assessment methods, including psychophysiological measures. In fact, astonishingly few trials have utilized such measures despite the practical and theoretical importance of changes in the autonomic responses involved in phobic and psychosomatic disorders. One possible reason for this is that to be used appropriately as an assessment of treatment-produced change, physiological response should be measured in the presence of an appropriate phobic stimulus, a condition not easily reconciled with the demands of polygraphic recording. One of the best solutions to this may be that adopted by Paul (1966) in using relatively simple measures immediately before contact with an anticipated phobic situation, an approach which turned out to be a productive one.

3. The decade that followed Wolpe's comprehensive presentation of reciprocal inhibition as the principle underlying desensitization saw widespread application and acceptance of his ideas, but little or no work designed to directly test the underlying assumptions of the reciprocal inhibition model as applied to systematic desensitization treatment. As formulated by Wolpe, this includes the hypotheses that the imagination of phobic stimuli leads to an autonomic and muscular response similar in type to that produced by real phobic stimuli, that relaxation training is a relatively efficient method of reducing or preventing these responses, and that this process of inhibition has some lasting effect extending into and influencing autonomic responses in real-life phobic situations.

Investigations into the mechanism of desensitization have primarily adopted the strategy of analyzing the treatment package into its component parts and assessing the relative contribution of each part singly or in combination (Rachman, 1967). While these investigations have produced some evidence that desensitization is most effective when both relaxation and the presentation of graded phobic images are combined, they do not bear directly on the hypotheses outlined above. To do this it would be necessary to design an experiment examining directly the immediate physiological and psychological effects of contiguous and noncontiguous presentations of phobic stimuli and relaxation. While such a crucial experiment is still lacking, there has been an increasing interest in the critical examination of these hypotheses using psychophysiological recordings during relaxation and the imagination of phobic stimuli, which allow a better evaluation of the present status of the reciprocal inhibition theory.

4. Also of relevance to the postulated underlying mechanisms of desensitization are a handful of analogue studies in which a process, held to resemble desensitization in its mode of action, is demonstrated and assessed using psychophysiological measures. Such analogue studies cannot, of course, prove anything about what happens in desensitization treatment, but they do indicate that a number of alternative models are available which may adequately describe the treatment process; for example, reciprocal inhibition, extinction/habituation, and cognitive rehearsal models.

UNCONTROLLED CASE STUDIES OF PATIENTS UNDERGOING DESENSITIZATION TREATMENT

Studies classified under this heading are invariably characterized by the application of only one autonomic variable, mainly, skin resistance (galvanic skin response—GSR). This choice may be determined both by the ready availability of portable equipment and by widespread assumption that skin resistance is superior to all other measures in sensitivity of response to emotion-provoking stimuli. This assumption is frequently made despite the considerable work on relative specificities in autonomic responding (Lacey, 1967) and the general criticisms of expecting simple relationships between physiological and psychological variables referred to earlier.

Single case studies of this type have been reported by Meyer (1957), Clark (1963), and Paul (1969b), all of whom assume that the presence or absence of a GSR to an imaginal item constitutes evidence of the degree of experienced anxiety. Similarly, Hoenig and Reed (1966), Seager and Brown (1967), and Gale, Hyman, and Ayer (1970) gave unsystematic reports of the clinical use of skin resistance in desensitization. While all these studies seem based on the same assumption, Hoenig and Reed expressed some skepticism based on the apparent dissociation observed in individual patients between GSRs to phobic words, images, or real objects, and verbal report. Despite this, general agreement seems to exist that a steady increase in skin resistance accompanies relaxation, while the responses to imaginal items decrease with repetition.

A somewhat more systematic approach was used by McAllister (1966)[3] who found that in some patients the amplitude of the GSR to phobic images was significantly greater than that to neutral images before desensitization, but that there was no difference between images at the end of treatment due to the usual decrement in response to phobic images.

The most informative of the uncontrolled studies have been reported by Agras (1965, 1967). In the earlier report, Agras described the spontaneous recovery of subjective anxiety to imaginal items following desensitization sessions in six patients said to have multiple phobias. Fifty percent of the desensitized items showed some degree of recovery in subjective anxiety when represented in a subsequent session, and this effect was paralleled by a similar recovery in skin resistance response to those items. In the later study, Agras reported the relationship between progress through the imaginal hierarchy, skin resistance response to test items and progress in real life, in five agoraphobic patients. The relevant measures of treatment progress were derived from a cumulative record of imaginal items completed, that is, imagined without reported anxiety, and a detailed self-report elicited from the patient of all traveling that was successfully carried out between treatment sessions. Prior to specified treatment sessions, five preselected items spread over the patient's hierarchy were presented in imagination, and the mean GSR amplitude calculated. Graphic presentation of the obtained data supports the author's claim that cumulative progress through the hierarchy was closely correlated with diminished GSRs to imaginal test items in three of the five patients, while progress in real-life practice tended to lag behind. In the other two patients, while hierarchy progress still preceded change in practice, skin resistance responses to the test items changed last of all.

---

[3] McAllister, H. The physiological correlates of anxiety. Paper presented at the meeting of the British Psychological Society, London, December 1966.

Wolpe and Flood (1970) reported a study of the GSRs to imaginal items at different hierarchy levels in five clients trained in relaxation, and in five who had not received such training. While a control (nonrelaxation) group was tested, the results are included in this section, as numbers were too small for statistical analysis. Subjects were given four desensitization sessions in each of which a complete hierarchy of five items was presented once, in ascending order. The results are presented graphically and show a fairly clear decline in size of GSR over sessions in the relaxed subjects, but little or no change in the unrelaxed subjects. However, while responses in the relaxation group tended to be greater with higher hierarchy items, as might be expected, the rate of response decrement was also greater with the higher items, contrary to prediction. Also unexpected was the tendency (attributed to chance) for the responses of nonrelaxed subjects to be less than those of the relaxed group on first presentation.

While it is not possible to draw any firm conclusions from the studies reported in this section, there is sufficient accord on some issues to justify the following statements.

Desensitization usually results in a reduced degree of skin resistance response to imaginal phobic stimuli, although it is unclear whether this should be described as a simple habituation to repeated stimuli, or attributed to the action of reciprocal inhibition. While most authors have found the application of skin resistance measures clinically useful, a number of cases were reported in which an apparent dissociation occurred between skin resistance responses and other behavioral changes, thus rendering the general clinical usefulness of the GSR as a single measure extremely doubtful.

Desensitization Trials Using Physiological Assessment

This section considers those controlled studies which have the primary aim of establishing the effectiveness of some variant of the desensitization procedure, but which have incorporated physiological measurement as part of an assessment battery.

Twelve asthmatic patients, 6 adults and 6 children, were treated by Moore (1965) in a controlled comparison of desensitization, relaxation plus direct suggestion, and relaxation alone. Assessment methods included a self-report of the number of attacks and estimations of respiratory maximum peak flow at the beginning and end of each session. Mean change in the number of self-reported asthmatic attacks tended to favor desensitization, but all treatments produced subjective improvement, and no significant differences were found on analysis of variance. Change in maximum peak flow, however, showed little effect due to treatments other than desensitization, which produced a marked improvement significantly greater than relaxation or suggestion.

The most thoroughly controlled trial in which physiological measures were used to assess outcome is that of Paul (1966). Students with fears of public speaking were allocated to desensitization, psychotherapy, attention/placebo, or no-treatment groups, and assessed before and after the treatment using self-report, behavioral, and autonomic measures. Each subject was required to give a short speech before a small audience, pulse rate and palmar sweat being measured 90 seconds and 30 seconds, respectively, prior to the speech. Self-ratings and behavioral signs of stress both responded more to desensitization than to the other treatments, but both the insight and the placebo treatments showed marked improvements in these measures. This was less true, however, in the case of physiological assessment, where significant change was confined to the desensitization group.

As might have been expected from previous reports of the relative independence between different parameters of fear, relationships between change scores derived from different assessment procedures tended to be fairly low. This was most obvious in the case of the correlations between the extent of change in physiological measures and in self-ratings of anxiety, which were extremely low and nonsignificant. The physiological assessment was related to the degree of behavioral change, however, and this latter measure correlated

.29 with pulse rate change, .25 with palmar sweat change, and .35 with a composite physiological score derived from both measures. Paul concluded from these results that desensitization does significantly reduce physiological activity in subjects anticipating a real-life phobic situation, that of speaking in public, while this effect was both less marked and less consistent following other treatments. In addition, an association, although a low one, was found between the extent of this physiological change and the improvement in the subject's performance under phobic stress.

Skin resistance levels during the presentation of real-life phobic stimuli (snakes) were used by Lomont and Edwards (1967) as one measure of outcome in a comparison of desensitization with and without relaxation. On retest after treatment both groups showed a large reduction in skin resistance response, and there was no difference between treatments with this measure, although behavioral and subjective changes showed a significant superiority for desensitization with relaxation.

Both Zeisset (1968) and Leitenberg, Agras, and Barlow (1969) recorded skin resistance responses during treatment sessions and found no relationship between the outcome of desensitization and the frequency of elicited GSRs. In commenting on the finding that some item presentations elicited a report of anxiety in the absence of a detectable GSR, Leitenberg et al. suggested that physiological anxiety responses might extinguish more rapidly than the verbal report of anxiety.

Reduction in the GSR to both real and imaginary phobic situations before and after desensitization were studied by Barlow, Leitenberg, Agras, and Wincze (1969). Following imaginal desensitization, GSRs to phobic imagery declined significantly, while responses to the real situation showed little change. In vivo desensitization, on the other hand, produced significant reductions in GSR to both test situations, suggesting good transfer back from real to imagined scenes, but little or no forward generalization from the imagined to the real situation. The same workers have reported a comparison between a "shaping" procedure in the presence of the real phobic object, and imaginal desensitization (Barlow, Agras, Leitenberg, & Wincze, 1970) with very similar findings.

A relatively novel psychophysiological measure, the penile plethysmograph, was used by Bancroft (in press) to contrast the effects of desensitization and aversion therapy in the treatment of homosexual patients. In general both methods produced a very similar amount of change as determined by a series of attitudinal and behavioral measures, and both treatments were associated with significant increases in erectile response to heterosexual pictures. As might be expected in view of the results reported earlier, the extent of erectile change did not correlate significantly with improvement measured in other ways.

In a comparison between desensitization and flooding (implosion) treatment, Boulougouris, Marks, and Marset (in press) recorded skin resistance and heart rate responses to the description and imagination of phobic scenes before and after each treatment. The results showed significant superiority for flooding over desensitization in terms of ratings made by the therapist and by an independent assessor, and in terms of the diminished physiological responses to the imagined test items. Changes in the physiological response to the fantasy test items were generally insignificant after six sessions of desensitization in these patients, but six sessions of flooding produced significant reductions in both heart rate and skin conductance response during phobic fantasy.

In general, it is possible to arrive at somewhat more definite conclusions from the survey of controlled studies than from the previous section. Moore's study alone indicates the necessity for including appropriate physiological measurements in the assessment of the effects of desensitization especially where somatic stress responses are prominent features. In Moore's and in Paul's study, it was convincingly demonstrated that desensitization was responsible for a significant diminution in the physiological response to stress, and that other procedures involving relaxation, suggestion, and a therapeutic relationship, tended to produce mainly subjective improvements and relatively little physiological change. It seems very likely that while many types of psychological procedures can influence verbal behavior, desensitization alone of those mentioned, is equally effective in reducing the autonomic components of fear. However, it may be that other procedures, such as flooding, will be found as or more effective in reducing autonomic responses to phobic stimuli.

In all the studies cited, there is evidence of a considerable degree of dissociation between physiological, behavioral, and subjective changes during treatment. Although all systems tend to change in the same direction, typically there is little correspondence between the amount of change in different systems. One possible explanation has been advanced by Leitenberg et al. (1969), that autonomic responses extinguish at a different rate than verbal reports of anxiety. This may be true of the GSR, which is notorious for its rapid habituation, without necessarily being true of other autonomic responses. The GSR also shows evidence of a marked degree of specificity in response decrement to different phobic stimuli. Cessation of GSRs to imaginal stimuli does not necessarily effect responses to real-life stimuli, although generalization in the reverse direction may occur (Barlow et al., 1969).

Psychophysiological Analysis of the Components of Desensitization

Analysis of desensitization into its component parts has been usually accomplished by a division into the specific effects of relaxation training and of the presentation of graded phobic stimuli (Rachman, 1967). For the purpose of this discussion, consideration of the possible effects of nonspecific components such as suggestion, positive expectation, and encouragement from the therapist (e.g., Leitenberg et al., 1969) are not included, since these have been discussed in detail by Wilkins (in press).

Relaxation

Three methods of training in relaxation are commonly used in desensitization, Jacobson's method of progressive relaxation (Jacobson, 1938), the direct induction of relaxation by hypnosis, and autogenic training (Schultz & Luthe, 1959). Autogenic training has been used less frequently in combination with desensitization, but has been described by Kondas (1967) and Moore (1965), and appears to produce satisfactory results. It is therefore assumed here that the active agent contributing to the importance of relaxation in desensitization is common to all three methods, and for this reason studies examining the physiological effects of each are described.

The physiological effects of hypnotically induced relaxation have been extensively studied, but very infrequently in the form of controlled experiments where the effects of hypnotic suggestion are contrasted with self-induced relaxation. Barber and Hahn (1963) reported finding no physiological effects during initial hypnotic suggestions of relaxation, which were not also seen in a control group simply asked to sit quietly.

Repeated use of autosuggestion, as in autogenic training, has usually been reported in the form of uncontrolled case studies (e.g., Schultz & Luthe, 1959), but Geissman and Noel (1961) found a number of physiological changes characteristic of autogenic relaxation that were not present in subjects at rest. These included decreased respiration rate, absence of GSRs, and electroencephalogram (EEG) with predominating alpha rhythm or occasional theta discharges. These results would seem reliable in that they have been replicated by several other workers (e.g., Jus & Jus, 1969), and it is perhaps surprising that more rigorously controlled studies have not been performed.

The most commonly cited evidence for the physiological effects of relaxation in desensitization derives from the work of Jacobson (1938). Using extremely lengthy training in muscular relaxation, Jacobson (1934) found decreases of muscle activity (electromyogram —EMG) in trained patients, but not in untrained controls. In two further studies (1939, 1940), both blood pressure and heart rate were found to drop during training sessions. Similar effects were not found when patients were asked to take daily rests for a year (1942). Unfortunately, Jacobson's results are of little scientific value due to his neglect of statistical procedures and of appropriate control groups. None of the controls used had regular contact with the therapist or the testing environment, as did the trained patients. It is quite possible to account for all the reported results in terms of either adaptation effects or the spontaneous remission of symptoms over time.

The first controlled study of brief muscular relaxation as used in desensitization was carried out by Grossberg (1965). Thirty male students were divided into three groups, one of which received recorded instructions in relaxation for two 40-minute periods, approximately 5 days apart, while one control group filled the same time by listening to self-selected relaxing music, and the other attempted self-relaxation without further instructions. Electromyograms from the forehead and forearm were recorded throughout, as was skin resistance and heart rate. No differences in general level were found among the groups in any of the measures used, indicating that the relaxation training procedure did not have any marked physiological effect.

A rather similar study was reported by Paul (1969c) in which 60 female students were divided into three groups according to whether they received training in muscular relaxation, hypnotic relaxation, or attempted self-relaxation without instruction. Electromyograms from the forearm, heart rate, respiration rate, and skin resistance were measured continuously through the two ½-hour sessions. Compared with the self-relaxation group, the group trained in muscular relaxation was found to have a significantly greater decrease of physiological activity over the sessions in all measures except skin resistance, where no significant differences were obtained, and the trend of the results was in the opposite direction.

It is difficult to reconcile the very different results of Paul's and Grossberg's studies in view of the rather similar procedures used; in addition, there seems some contradiction between the results of hypnosis obtained by Paul (1969c) and by Barber and Hahn (1963). Further negative results, obtained incidentally from more recent studies in which students have received similar training in relaxation, tend to support Grossberg's conclusion that very brief training is only slightly, if at all, effective in reducing the level of autonomic activity when compared with the resting state (Lehrer, 1969;[4] Van Egeren, Feather, & Hein, in press).

A somewhat different procedure was adopted by Mathews and Gelder (1969) in a study of the effect of relaxation training in phobic patients. Fourteen patients were given two training sessions derived from Jacobson's abbreviated technique (1938) and were then tested during a further relaxation session and a control period, while forehead EMG, heart rate, respiration rate, and skin resistance were continuously recorded. Significant reductions in skin conductance level and fluctuation rate were found during relaxation as compared with the control condition. Electromyograms remained consistently lower throughout the relaxation session, and both heart rate and respiration rate tended to drop more during relaxation, but this effect did not achieve an acceptable level of significance. All measures, especially skin conductance fluctuations, produced significant correlations with subjectively estimated depth of relaxation, with the exception of EMG which showed no such relation.

The main differences between this and previous studies, apart from the use of psychiatric patients, is that the subjects were relatively better practiced and familiar with relaxation, so that the active muscular exercises used earlier could be discarded, and the procedure used was similar to that employed in a standard desensitization session. The most striking contrast between the results of this study and those of Grossberg and Paul, that is, the skin resistance findings, may most probably be attributed to this difference in procedure.

It is difficult to summarize the available knowledge concerning the effects of brief relaxation training, and the contradictions in the results suggest the operation of at least some factors not yet fully understood. Only Paul has produced any evidence that the particular procedure adopted had any specific effect related to the method of training, independent of the general effects of adaptation and the reduction of external stimulation. It seems reasonable to suggest that several specific procedures may have the effect of reducing autonomic and muscular activity, but that whether these effects are observed, and in which measure, depends on details of procedure and timing, the length of training, and the nature of the comparison control procedure.

---

[4] Lehrer, P. M. A laboratory analog of desensitization: Psychophysiological effects of relaxation. Paper presented at the meeting of the Society for Psychophysiological Research, Monterey, California, October 1969.

If the results obtained by Paul are accepted in contrast to those of Grossberg, the use of active exercises in training may have the effect of inducing transient periods in which muscular, cardiac, and respiratory activity is minimal. Relaxation as used in actual desensitization, however, may be associated with a marked diminution in skin resistance activity, which is correlated with sensations of subjective relaxation and calm. Further periods of practice, as are employed in autogenic training, may be associated with even more profound changes, observable in the EEG record.

Phobic Stimuli

The relationship of autonomic responding to the intensity of experienced emotion, in particular anxiety, has been the source of a long continued controversy. A review by McCurdy (1950) is commonly quoted in support of the use of the amplitude of the GSR as an index of emotional intensity, since all the relevant studies found by McCurdy reported positive, and usually high, correlations between GSR and judged intensity.

Although the studies reviewed by McCurdy generally refer to relatively physical stimuli, such as emotional words or electric shocks, it is fairly well-established that the response to imaginal or vicariously experienced stimuli produce results similar to directly experienced stimuli (Barber, 1965; Craig, 1968; Miller, 1951).

Lomont and Edwards (1967) and Barlow et al. (1969, 1970) have reported the only controlled studies which included physiological measurement of the response to real-life phobic stimuli, in this case, the GSR in snake phobic subjects to presentations of a snake at varying distances. Since no control stimuli were presented at the time, the only conclusion possible from these studies is that the GSR to phobic stimuli declines following desensitization or exposure to the real object.

An alternative method of examining the physiological effects of phobic stimuli is to compare responses to viewing pictures of neutral and fear-evoking objects. Geer (1966) assessed the change in skin resistance caused by viewing pictures of spiders and pictures of snakes, in subjects who had expressed a fear of spiders only. Subjects reporting an intense fear had larger responses to pictures of spiders than those who reported mild fear, while similar groups did not differ in responding less to pictures of snakes. The difference between the first group and the other three was highly significant on first showing of the spider slide ($p = < .02$), less so on second showing ($p = < .05$), and just failed to reach significance on the third and final presentation. While clear-cut differences in the predicted direction were found, it is of interest that habituation effects rapidly and effectively reduced these differences to the point of nonsignificance.

Studies essentially similar to that of Geer have been reported by Wilson (1966, 1967). In the first of these experiments Wilson established that skin resistance responses to a series of slides, including pictures of landscapes, snakes, and spiders were not significantly effected by the picture content when shown to subjects presumed to be without abnormal fears. Two phobic subjects were tested individually, one of whom complained of a fear of cats and the other of spiders. Responses to pictures of cats in the first subject were found to be considerably greater than responses to 11 other animal pictures on four successive trials, while a comparison of alternating spider pictures and landscapes in the second patient showed a significantly greater response to spiders, which habituated to the point of no difference over four sessions.

In the second study, 10 students who had expressed a fear of spiders were contrasted with 10 control subjects without such fears. Eight pictures of spiders were projected in alternation with eight landscapes while the GSR was measured as before. Use of the ratio of response to spider pictures, divided by the response to landscapes, gave perfect separation between phobic and nonphobic subjects. Once again, however, a clear tendency was apparent for responses following phobic slides to habituate, although this took longer to complete than did responses to neutral pictures. As in the study by Geer, no additional measures were used that would allow conclusions as to the subjective state experienced by the subjects, or even generalization to other physiological variables. In the absence of self-report data, it would be obviously fallacious to assume that the large GSRs to phobic pictures reflected phobic anxiety; they may equally reflect more careful attention paid to them by subjects having a special interest (Flanagan, 1967).

The stimuli most commonly associated with desensitization are produced in fantasy, as the use of imagery makes up in flexibility what it lacks in objectivity and accessibility to precise measurement.

A very early study, which is relevant to this issue, was carried out well before the use of desensitization was established (Rowland, 1936). Four normal subjects, capable of a deep hypnotic trance, were given 12 scenes to visualize, which varied in their "exciting" quality. Although the use of few subjects and primitive physiological recording techniques limit generalization, the study was carefully done, and the conclusions may well be considered useful. Pulse rate, respiration, and skin resistance were all found to relate to the prejudged scene intensity, while repeated presentations of the scenes led to a diminution in these responses.

In McAllister's study (see Footnote 3), 10 phobic situations drawn at random from each patient's hierarchy, preserved in hierarchal order, were interspaced with neutral scenes and presented for imagination while skin resistance was recorded. On initial presentation, five out of eight patients responded significantly more to the phobic than to the neutral stimuli.

The effect of imagining phobic stimuli has also been studied using plethysmographic measurement of forearm muscle blood flow (Gelder & Mathews, 1968). Three imaginary situations were found for each of 19 phobic patients, a neutral scene and two phobic scenes, one of which was taken from the bottom of a desensitization hierarchy and one from the top. The patients were asked to imagine each scene in turn and to rate immediately afterwards any subjective anxiety experienced during the image. Mean blood flow was raised during the imagination of both phobic scenes, although only during the higher phobic scene was the level reached significantly greater than that during neutral scenes.

A more thoroughly controlled study was reported by Grossberg and Wilson (1968), in which forehead EMG, heart rate, and skin resistance were recorded as subjects imagined fearful and neutral scenes. Substantial increases in heart rate and skin conductance were observed during the reading of both phobic and neutral scenes, but significant differences between scenes in heart rate and skin conductance responses were confined to the imagination period. However, by reanalyzing in terms of the frequency of responses, irrespective of amount, reading of fear scenes was found to produce significantly more skin conductance increases, and imagining produced significantly more heart rate increases. Repetition of these scenes showed a systematic decline in responses, closely resembling a typical extinction or habituation curve. Electromyogram data did not show the regularity apparent in the autonomic variables, and no significant differences were found between the scenes in terms of muscle activity.

A further group of control subjects, not reporting any of the relevant fears, was matched with experimental group subjects and given identical instructions. Only heart rate response during the reading of fearful scenes showed a significantly greater effect in experimental than in control subjects, while the difference in heart rate and skin conductance response found earlier between fearful and neutral scenes was still present in the control subjects. Since the two scenes should not have been emotionally different for control subjects, this result is against prediction and suggests that in the presentation of the fearful scenes the presence of some cues had a relatively greater effect even in nonphobic subjects. Finally, correlations were calculated between physiological responses and subjective estimates of vividness. Only one correlation was significant, that between heart rate response to imagery and vividness ($r = .60$).

It can reasonably be concluded that predictable autonomic responses follow the imagination of fearful scenes, such as those used in desensitization, and that these responses decline systematically with repetition, even in the absence of relaxation training. The persistence of scene differences in control subjects, however, suggests strongly that voice cues, subject expectancies, or some other unknown factors are involved in producing these differences.

The effect of imagining fearful and neutral scenes on respiration changes in 55 students was reported by Rimm and Bottrell (1969). As in the previous experiments cited, significant differences emerged in physiological response to the different scenes; increase in respiration rate, but not depth, was significantly greater during the fearful scenes. No significant decline in response was found when each scene was presented once more. Prior to actually imagining the scenes, each subject had given a general rating of his ability to visualize scenes clearly, and a correlation of .3 ($p < .05$) was found between this rating and respiratory response to the fearful scene.

The nature of the interaction between the effects of phobic imagery and of relaxation training has been further studied by Paul (1969a). Sixty female students were allocated to three groups receiving, respectively, training in muscular relaxation, hypnotic suggestions of relaxation, and no special instructions, as was described earlier (Paul, 1969c). Before the two training sessions, the subjects were asked to imagine items previously picked as not disturbing (neutral) and as disturbing (fearful) from the Wolpe-Lang Schedule, in that order.

After the training sessions had been completed, all subjects were again asked to imagine the fearful scene, first for 5 seconds, then for 10, and finally for 20 seconds, only the last image being scored. As before, forearm EMG, heart rate, respiration rate, and skin resistance were measured, but analysis was confined to those particular measures in which an individual had shown signs of responding on initial presentation of the fearful scene. Both the muscular relaxation group and the hypnosis group showed a significantly greater decline in physiological response than the control group, and the extent of this decline was found to be related to the average reduction in physiological activity during relaxation ($r = .46$).

In addition, and contrary to the results found by Grossberg and Wilson, Paul (1969a) reported that control (untrained) subjects actually showed an increase in composite response to the fearful image on the final presentation. This result raises the possibility that some unknown procedural difference between the methods used by Paul (1969a) and Grossberg and Wilson (1968) seriously influenced the experimental results. One possibility is that Paul's control group, unlike his other two groups, had not habituated to instructions from the therapist during training, and may well have overreacted to the imagination instructions following a prolonged period of silence. Despite this, it may at least be concluded that repetition of fearful imagery in the relaxed subject leads to a diminution of autonomic response, although it remains unclear whether relaxation might produce an immediate reduction in the response to initial image presentation, or whether relaxation is associated with a more rapid rate of decline in responding.

One of the most extensive studies of the physiological effects of phobic imagery has recently been completed by Lang, Melamed, and Hart (1970). In the course of a comparison between automated and live desensitization, which showed the automated treatment to be at least as effective as conventional methods, heart rate, respiration rate, forearm EMG, and skin resistance were measured throughout treatment. Electromyogram activity was generally found to be minimal and was discarded. An examination of the physiological events accompanying imagined scenes that provoked an anxiety signal from the subject showed that heart rate, respiration rate, and skin conductance all tended to peak on that item. Subsequent re-presentations without anxiety signals were marked by a consistent decline in heart rate response, and although the other measures also showed a marked decline, this was less obviously monotonic.

The most striking findings were of high correlations between these heart rate responses and a composite measure of the behavioral and self-reported reductions in fear at the end of treatment. Subjects who showed the highest heart rate during a signaled item also responded most to treatment ($r = .75$; $p < .05$), while the correlation between the decline in this heart rate response to repeated presentations (habituation) and good outcome was even higher ($r = .91$; $p < .01$). The rate of heart rate decline at other points in the treatment (e.g., within relaxation sessions) did not correlate with outcome, nor did skin conductance or respiration responses appear to relate to treatment effectiveness.

Lang concluded from this first experiment that subjects who are more responsive autonomically, as assessed by heart rate during phobic scenes, and in whom this physiological change is synchronous with self-reported anxiety, showed both a more clear-cut decline in autonomic response and subsequent better behavioral response to treatment. The subjects who responded poorly to the treatment gave in general a similar number of anxiety signals, but these were not synchronous with physiological response.

In a second experiment, responses to phobic scenes in students having a fear of public speaking were contrasted with those in students having a fear of spiders. Five preselected hierarchy items were presented in random order and interspaced with neutral items, both subjective anxiety and vividness of imagery being rated on a 0–4 scale immediately after each item. Heart rate consistently increased as higher hierarchy items were presented, and similar although somewhat less consistent results were found for skin resistance and respiration. Correlations between physiological response and self-report of anxiety or vividness tended to be maximal when heart rate was used, although a high correlation was obtained between skin resistance response and reported vividness.

A complex, but well-controlled study has been recently reported by Van Egeren and his colleagues (Van Egeren et al., in press), in which imagined items spaced over a hierarchy of public speaking situations were presented in balanced order to 15 subjects trained in relaxation and 15 untrained subjects. Heart rate, respiration rate, skin resistance, and digital pulse amplitude were recorded throughout, and all of these measures significantly differentiated between phobic and neutral items. The effects of item repetition, and the interaction with relaxation and hierarchy position were unexpected, and in some respects inconsistent with earlier work. Only digital pulse amplitude response consistently declined with item repetition, regardless of training in relaxation, and this decline did not show any tendency to recover between sessions (cf. Agras, 1965; Grossberg & Wilson, 1968). However, there was an overall tendency for skin resistance responses to all items to be less marked in the relaxation group. Items higher in the hierarchy (i.e., more intense) elicited significantly greater skin resistance and digital pulse amplitude responses, but contrary to expectation the rate of response decrement was greater for these higher items than for those low in the hierarchy. Again, contrary to expectation, where relaxation training had significant effects, they were most marked for the scenes of highest intensity. While the results suggest that the response decrement to imaginal items is not the predicted simple function of repetitions, relaxation training, and hierarchy position, some serious reservations must be made before making a direct comparison with desensitization treatment. For example, items were not given in ascending order, and the same item was not repeated consecutively until anxiety was reduced. In fact there was no significant decline in anxiety ratings throughout the experiment (two sessions), nor were there any significant differences in autonomic resting levels due to relaxation, suggesting that the experimental procedure was sufficiently different from desensitization treatment to severely limit the generality of any conclusions drawn.

A considerable amount of evidence has been summarized in this section, and the conclusions to be derived vary from the fairly certain to the wholly hypothetical. In study after study, phobic stimuli have been shown to evoke significantly greater autonomic responses than neutral stimuli. It remains unclear, however, to what extent these autonomic responses are entirely attributable to phobic anxiety, and to what extent factors such as expectation or cues given by the therapist may also be involved. Johnson and Campos (1967) have implicated the subject's knowledge that he is required to make a subsequent verbal report on his experience during imagery, as a critical factor in determining the extent of heart rate responses to neutral images. However, the finding of significant relationships between heart rate and anxiety or vividness in two independent studies goes at least some way to support the former explanation, and is consistent with Wolpe's account of the mechanism of desensitization. Both heart rate and other cardiovascular measures, forearm blood flow and pulse amplitude, have been found to show increased responses as items higher in the hierarchy are presented. Heart rate may be a particularly responsive measure during phobic imagery, as opposed to real-life visual stimuli, because of the tendency described by Lacey (1967) for heart rate to decelerate during orientation reactions to external stimuli and to accelerate during mental tasks when stimuli from the external evironment are ignored.

When phobic scenes are described, prior to imagination, skin resistance responses may be more frequent than heart rate responses (Grossberg & Wilson, 1968), and when pictorial stimuli are used, the GSR consistently differentiates between phobic and neutral stimuli. It seems, therefore, that while autonomic responses reliably discriminate phobic from neutral stimuli, the particular measure most appropriate may depend on the type of stimuli presented.

Some researchers, with the notable exception of Paul (1969a), have reported that autonomic responses to phobic stimuli show a marked tendency to diminish with repetition, whether or not the subjects have been trained in relaxation. Paul's study was the only one in which responsiveness appeared to increase with repetition, although the usual reduction occurred in subjects trained in muscular or hypnotic relaxation. The findings of Van Egeren's experiment, although difficult to interpret, suggest that the presence and extent of such a decrement is a complex function of the particular physiological measure used, interacting with other interpolated stimuli, the presence of relaxation, and hierarchy level of the items presented. However, as was pointed out earlier, the procedure involved in this experiment differed in many respects from desensitization treatment.

Lang's intriguing results provide evidence that the subject's facility in obtaining imagery vivid enough to elicit clear autonomic responses synchronous with the verbal report of anxiety, as well as the subsequent diminution of such responses, are critical factors associated with the successful outcome of treatment. It is, of course, possible to argue that this finding reflects only the incidental side effects of a cognitive change occurring in desensitization; however, the stronger hypothesis would seem to be that the presence of autonomic responses to imagery, the decrement of such responses with repetition, and possibly the subject's awareness of these changes may all be essential parts of the desensitization process.

Analogues of Desensitization

Studies included in this section are all primarily concerned with investigating or demonstrating mechanisms which may underlie desensitization, and all arrive at different models of the treatment process, based on experimental analogues. The best known theoretical account of desensitization, that of reciprocal inhibition, or counterconditioning (Wolpe, 1958), has been demonstrated in analogue form by Grings and Uno (1968). The effects of visual cues that had been associated with muscle relaxation and electric shock were compared singly and in combination. Both skin resistance and pulse volume responses tended to be greatest to the shock cue alone, were slightly reduced to a combination of shock and neutral cues, and were least to the shock and relaxation cue combination.

These results suggest strongly that conditioned autonomic fear responses and learned relaxation can reciprocally inhibit one another and thus provide a useful experimental situation in which variables relevant to desensitization may be manipulated. However, further work would be required before it could be concluded that such counterconditioning actually occurs in the treatment situation, or that this process is a necessary part of successful desensitization.

An alternative to the counterconditioning model has been recently proposed by Lader and Mathews (1968) based in part on the findings of a relationship between habituation rate to simple stimuli and the outcome of desensitization (Lader, Gelder, & Marks, 1967). In this latter study phobic patients were tested prior to desensitization in order to assess rate of habituation of the GSR to repeated auditory stimuli. Behavioral signs of anxiety rated during pretreatment testing were significantly associated with delay in habituation ($r = .36$) and with the frequency of skin conductance fluctuations ($r = .61$). Clinical assessments of improvement after desensitization were also correlated with pretreatment physiological measures to a significant degree; .49 in the case of habituation rate and − .42 in the case of skin conductance fluctuations. Although these results are explicable in terms of the expected relationship between high levels of anxiety and both a poor response to desensitization and autonomic reactivity, it is possible to postulate a more direct link between habituation and desensitization. Lader et al. have proposed that desensitization may be viewed as a process similar to that of habituation taking place under optimal conditions, that is, while the patient remains at the lowest level of arousal consistent with consciousness.

As in the previously cited study these results may only be considered suggestive and do not provide direct evidence as to the physiological processes occurring in desensitization. Attempts at providing more direct evidence concerning the nature of these responses have been made by Van Egeren (in press) and Melamed (cited by Lang, 1970). Van Egeren concluded, on the basis of the studies previously discussed, that while any diminution in responses to imaginal stimuli is not effected by relaxation in the way predicted by counterconditioning, neither are the parameters of such a response decrement identical with those found in the case of simple physical stimuli (habituation). This would appear to be a very reasonable conclusion, but requires confirmation by further studies of the response decrement occurring when imaginal scenes are repeated without interpolated stimuli conceivably having disruptive effects.

Melamed recorded skin resistance responses in snake phobic students to both a series of simple auditory stimuli and a series of films showing snakes. Subjects were divided into three groups on the basis of preliminary instructions to simply attend to the films, to experience the full force of their emotional response, or to utilize training in relaxation. Both the "attend" and "relaxation" groups showed a response decrement to the phobic stimuli which correlated highly with habituation rate to auditory stimuli, and significant correlations were found between this habituation rate and a questionnaire measure of fear change, consistent with the results obtained with patients by Lader et al. (1967). Curiously, however, correlations between autonomic response decrement to the films and reported reduction in fear did not reach a significant level, so that these results cannot be considered as unqualified support for the habituation model of desensitization.

Folkins, Lawson, Opton, and Lazarus (1968) have studied methods of reducing the effects of viewing a stressful film as an analogue of desensitization. The effects of prior relaxation, cognitive rehearsal with stressful imagery, and both procedures together were contrasted with no treatment in terms of their efficacy in reducing heart rate and skin conductance responses to a film portraying a fatal accident. Contrary to prediction, subjects who had received cognitive rehearsal only, that is, desensitization without relaxation, reported less anxiety during the film than any other group. Skin conductance showed a peak during the stressful accident scene in all groups, but once again the cognitive rehearsal group had a less marked peak than any other group. Heart rate responses did not follow the same pattern, and no significant differences were found among the treatment groups for this variable. Correlations between psychological and physiological measures of stress were generally low, the highest being that between self-reported stress and peak heart rate during the filmed accident ($r = .32$, $p < .01$).

One possible explanation suggested by Folkins to account for the relatively poor showing of desensitization in this experiment is that the tape-recorded imagery instructions proceeded rapidly and inflexibly to stressful scenes, and may thus have disrupted relaxation. Some weight can be attached to this suggestion, as a greater skin conductance response occurred to imagined stressful scenes in the desensitization group than in the cognitive rehearsal group. Thus the addition of relaxation instructions appears to have been associated with an increased autonomic response to imagery, contrary to reciprocal inhibition theory. Folkins considered that his results provide support for the role of cognitive processes in desensitization, and cited the criticisms of stimulus-response theorists made by Breger and McGaugh (1965). However, the considerable differences between this experiment and treatment practice, such as the use of nonphobic subjects and the anticipation of a stressful film in place of the usual phobic imagery, must limit the degree to which generalizations can be made to desensitization.

Although not employing physiological measures, consideration of an experiment by Valins and Ray (1967) may be relevant. In this study, snake phobic subjects listened to a recording which they believed to be of their own heart sounds and were thus persuaded that their heart rate decreased as they watched slides of snakes. In contrast to a control group, these subjects showed a significant reduction in snake avoidance behavior, a result interpretated by the authors as evidence of important cognitive factors in desensitization. While it seems likely that the subjects awareness of reductions in autonomic response to phobic stimuli may have an important role to play in desensitization, Valins and Ray's experiment is not relevant to the postulate that real autonomic changes are necessary in order that desensitization should be maximally effective. In this connection it is perhaps a pity that this study did not in fact include heart rate recordings in order to examine the possibility that false feedback may have induced real heart rate changes in the same direction.

Perhaps the only conclusion that can be derived from these studies is that advocates of counterconditioning, habituation, and cognitive models of desensitization can all point to some experimental support from studies utilizing psychophysiological measurement. All these models may be heuristically useful, but their relative value can be provisionally assessed by a consideration of their ability to account for the results described earlier.

DISCUSSION

The evidence presented in the previous sections strongly supports the idea that psychophysiological measurement is of potential use in studying both the process and outcome of desensitization.

In an earlier review of psychophysiological approaches to psychotherapy, Lacey (1959) came to a similar conclusion as regards the study of psychotherapeutic process, but was less optimistic as regards the assessment of outcome. It may well be that one of the most important aspects of behavior therapy is that by defining more precisely the target behaviors and the procedures involved in treatment, objective measurement of autonomic and behavioral events becomes both more possible and more useful. Examination of physiological responses during desensitization has already led to the confirmation of some predictions regarding the processes involved, and the disconfirmation of others. The choice of appropriate physiological measures, the nature of the response to phobic imagery, the interaction of this response with relaxation, the relationship between changes in the different components of fear, and the theoretical mechanisms underlying desensitization are each considered in the light of the available evidence.

Choice of Physiological Measure

Although useful results have been obtained using respiration and EMG measures, the most consistently significant effects have been observed using skin resistance or cardiovascular responses, particularly heart rate. Lacey (1967) and Graham and Clifton (1966) have suggested that heart rate deceleration accompanies the intake of information from the environment, while concentration on internal mental events, or rejection of the environment, leads to heart rate acceleration. Evidence consistent with this hypothesis has been reported by Craig (1968) who found that imagining a painful stimulus produced as much heart rate increase as the actual stimulus itself, while the vicarious stimulation of watching a fellow subject experiencing pain was not accompanied by any such increase in heart rate. Skin resistance response, on the other hand, was more sensitive to the vicarious than to the imaginal stimulus. It might therefore be expected that heart rate should be a relatively sensitive indicator of response to phobic images, while skin resistance might be more responsive to externally presented phobic stimuli.

Grossberg and Wilson (1968) found evidence that skin resistance responses were more frequent during scene descriptions, while heart rate responses were more frequent during the actual image. Estimated vividness of the image was significantly correlated with heart rate, but not with skin resistance.

Lang et al. (1970) found a similar relationship between vividness and both heart rate and skin resistance response, although only heart rate correlated significantly with subjective anxiety. In this latter experiment heart rate response was also more closely related to hierarchy level than with the other measures used. Skin resistance responses have not always been successful in the consistent separation of phobic from neutral images (e.g., McAllister, 1966, see Footnote 3), although when phobic stimuli are presented visually (Geer, 1966; Wilson 1966, 1967) the GSR was a highly effective discriminator. The well known tendency of the GSR to habituation, however, rapidly reduced this discriminative power as the stimuli was repeated. The choice of the most appropriate measure would therefore seem to depend on the nature of the stimuli used, as well as on individual response specificity. However, in general, heart rate has proved most useful both as a measure of response to phobic imagery and as an assessment of outcome (e.g., Paul, 1966).

The addition of at least skin resistance would seem advisable on general grounds, and especially when real phobic stimuli are presented (Barlow et al., 1969) or as a measure of success in relaxation (Mathews & Gelder, 1969).

Effects of Phobic Imagery

One of the most consistent findings to emerge from this review is that imagined phobic scenes as used in desensitization evoke larger autonomic responses than scenes judged to be neutral, and that with repetition these responses show a consistent decrement. Wolpe (1958) has commented that under certain circumstances, anxiety does not dissipate between images but appears to summate. This clinical observation has not generally been supported by psychophysiological evidence, with the exception of one experiment by Paul (1969a). Such a response increment would thus seem to be the exception rather than the rule, and its occurrence may depend on relatively high anxiety in the subject, or on the frequent application of high intensity stimulation. These conditions have been described as those which disrupt the usual process of habituation to repeated stimuli and may lead to an upward spiral of responses into panic (Lader & Mathews, 1968, 1970). They are also the same conditions prevailing in flooding sessions, suggesting that if continued for long periods this upward spiral must be self-limiting and in the end will also lead to decrement in autonomic anxiety responses (Boulougouris et al., in press). There is thus tentative evidence of two related procedures having a similar result; presentation of mild phobic images interspaced with a rest or relaxation period, or the presentation of intense phobic images under conditions of prolonged massed practice, both leading to a marked decrement of autonomic response to imagined phobic stimuli.

Effects of Relaxation

The evidence concerning the physiological effects of relaxation training suggests the possibility of transient reductions in cardiovascular, respiratory, and muscular activity, following exercises described by Jacobson (1938). Relaxation sessions following longer periods of training, however, may be accompanied by evidence of a generally lowered level of arousal (Mathews & Gelder, 1969). Despite this, no direct evidence has been found for one of the central postulates of reciprocal inhibition theory, that relaxation reduces or prevents the autonomic anxiety responses associated with phobic imagery. On the basis of Paul's (1969a) experiment, it may be tentatively concluded that a response to a phobic image, at least by the fourth presentation, is less in the relaxed than the nonrelaxed subject, without it being clear whether immediate inhibition, an accelerated rate of response decrement, or both, are involved. The effects of relaxation may, in fact, vary according to the hierarchy level of the imaginal item used, since Van Egeren et al. (in press) reported no effect with low hierarchy items, but a tendency for larger initial responses and more rapid decrement with repetition in relaxed subjects when high items are presented. Despite using fewer subjects, an almost identical result was reported by Wolpe and Flood (1970).

In the Folkins et al. (1968) analogue study, relaxation appeared to produce response facilitation rather than inhibition on the first presentation of a moderately intense image, while in a direct study of desensitization, Lang (et al. 1970) found evidence that a relatively large autonomic response to phobic images, followed by a rapid decay, is associated with a good treatment outcome. These results suggest the possibility that the rapid reduction of an initially large response is a critical factor in successful desensitization, and that relaxation may assist in bringing about this effect. Anecdotal evidence of the intense vividness of images experienced in a relaxed or drowsy state (hypnagogic imagery), and the recent work of Antrobus (1968) showing that the spontaneous occurrence of images increases as external information is reduced, both support the hypothesis that one of the effects of relaxation may be to increase the vividness of imagery experienced during desensitization. Correlations between subjective estimates of vividness and the degree of autonomic response (Grossberg & Wilson, 1968; Lang et al., 1970; Rimm & Bottrell, 1969) further suggest that more marked physiological responsiveness to phobic imagery may thus be produced. Hence, the evidence that has been reviewed gives rise to the hypothesis that relaxation may augment both vividness and the autonomic effects of imagery, while at the same time maximizing the response decrement with repeated presentations.

Relationship between the Components of Fear

Desensitization (and flooding procedures) have been found to have a relatively powerful effect in reducing autonomic anxiety responses. However, the dissociation between physiological, behavioral, and subjective changes noted in case studies of desensitization has been confirmed in controlled trials (Moore, 1965; Paul, 1966). Despite this, it seems extremely unlikely that no causal relationship exists between autonomic responding during desensitization and subsequent behavioral change. Indeed, Lang's most recent results (1970) suggest strongly that marked autonomic changes during treatment, synchronous with reported anxiety, are an essential feature of successful desensitization. An alternative explanation for the reported discrepancies is that causal relationships may be obscured by a tendency for the separate components of fear to change in differing amounts at different times (e.g., Lang & Lazovik, 1963).

Within imaginal desensitization, habituation of the GSR to phobic images may precede the elimination of anxiety signals (Leitenberg et al., 1969), and although this order may be reversed in some individuals (Agras, 1967), both of these usually precede change in actual phobic behavior. Barlow et al. (1969) have confirmed uncontrolled observations (Hoenig & Reed, 1966) that following imaginal desensitization, there may be relatively little reduction in GSR to the real phobic object, although the response to imagined scenes had greatly decreased. Behavioral change in this group was also small compared to a group who had been repeatedly exposed to the real phobic object, and this latter group showed a very marked decrement in GSR to further presentations of phobic objects; indeed, the mean response after treatment was less to the real than to the imagined scene. This important finding suggests a considerable degree of stimulus specificity in response decrement, and argues against the idea that elimination of autonomic responses to imaginal scenes generalizes to any great extent to the real-life situation. The various findings on temporal displacements in improvement can be heuristically described in terms of the following postulate. The most common order of change resulting from desensitization begins with a progressive reduction in autonomic response to imagery, associated with or followed by, decreasing reports of subjective anxiety. Behavioral change having the consequence of closer contact with the real phobic situation may then occur, followed by autonomic response decrement to the real-life stimulus and finally a reduction in self-ratings of anxiety to the phobic object or situation. Overlap between these processes must occur to some extent, and there may be a considerable degree of idiosyncratic response ordering in some patients. It is further postulated that such temporal displacements are responsible in part for the extremely low correlations observed between the extent of posttreatment change in the different systems, since testing during or shortly after treatment is likely to find subjects at different stages of progress.

Theoretical Mechanisms Underlying Desensitization

In view of the tentative conclusions that have emerged from this review, it is considered that each analogue described in the last section in fact represents an oversimplified account of the total desensitization process. It was earlier suggested that phobic imagery may be made more vivid and may evoke stronger initial autonomic responses when the subject is in a relaxed state. Should this be accepted it must clearly modify reciprocal inhibition theory to the extent of assigning both an initial facilitating and later inhibiting role to relaxation. If the habituation analogue is considered, then a possible explanation for the association of vivid imagery and large initial autonomic response with good outcome suggests itself. According to Sokolov's (1963) account of habituation, a neural model is generated by application of a repeated stimulus, which then acts to inhibit responses to incoming stimuli which match the internal model. In this way, observable responses gradually cease as the neural model is built up, unless the organism learns that occurrence of the stimulus leads to significant consequences. If a related process can be said to occur in imaginal desensitization, then it would clearly be necessary for vivid imagery to be experienced, otherwise construction of an internal model would be delayed or might not take place at all. On the other hand, Lader et al. (1967) have presented evidence that anxious patients with a high level of spontaneous autonomic activity habituated extremely slowly, and Lader and Mathews (1968) suggested that relaxation may be one way of counteracting this effect. Considering desensitization as analogous to habituation would therefore lead to the hypothesis that a state of low arousal, for instance, relaxation, produces two useful effects; it facilitates the occurrence of a vivid image and thus a better internal model, while at the same time it maximizes the rate of response decrement to that image.

The failure of generalization across systems or from imaginal to real stimuli presents further problems, however, There are several processes which may be involved in extending the effects of desensitization from decrement in response to imaginal scenes to a reduction in phobic avoidance behavior. Intuitively it seems likely that phobic avoidance is partially maintained by the anxiety associated with anticipation and mental rehearsal, and that following desensitization this anticipatory anxiety is reduced. Indeed, the pre- and posttreatment tests used by Paul (1966) to assess autonomic changes took place immediately before subjects were required to speak in public, so that there is some evidence of immediate posttreatment physiological change during anticipation of contact with a phobic situation. In this study, autonomic measurement was not continued throughout contact with the real-life situation, but the postulate put forward here would lead to the prediction that under these circumstances the effect of the imaginal desensitization would have been less marked, as was found by Barlow et al. (1969). Despite this, decreased autonomic responsiveness and lessened verbalization of anxiety during anticipation may well lead to an increased likelihood of approach behavior in the phobic situation.

Awareness of reduction in autonomic responses may lead to therapeutic attitude change as theorized by Valins and Ray (1967), while the expectations of both therapist and patient may provide a further motivating influence (Leitenberg et al., 1969). Once increasing contacts with a phobic situation are made in the absence of anticipatory anxiety, a decrement in autonomic responding to real phobic stimuli is likely to occur, leading finally to self-reports of no anxiety in real life.

To summarize, an attempt to encompass as much evidence as possible within a relatively simple account of desensitization leads to the postulate that the initial effect is one of response decrement in the autonomic and self-report components of anxiety during imagery, whether this is described as reciprocal inhibition, extinction, or habituation. Possibly aided by several different cognitive processes, this decrement leads to increased exposure to real phobic situations, and thus enables a similar decrement of autonomic and verbal responses to occur in real life.

REFERENCES

Agras, W. S. An investigation of the decrement of anxiety responses during systematic desensitization therapy. *Behaviour Research and Therapy,* 1965, **2**, 267–270.

Agras, W. S. Transfer during systematic desensitization therapy. *Behaviour Research and Therapy,* 1967, **5**, 193–270.

Antrobus, J. S. Information theory and stimulus-independent thought. *British Journal of Psychology,* 1968, **59**, 423–430.

Bancroft, J. H. The application of psychophysiological measures to the assessment and modification of sexual behaviour. *Behaviour Research and Therapy,* in press.

Barber, T. X. Physiological effects of "hypnotic suggestions": A critical review of recent research (1960–64). *Psychological Bulletin,* 1965, **64**, 201–222.

Barber, T. X., & Hahn, K. W. Hypnotic induction and "relaxation." *Archives of General Psychiatry,* 1963, **8**, 295–300.

Barlow, D. H., Leitenberg, H., Agras, W. S., & Wincze, J. P. The transfer gap in systematic desensitization: An analogue study. *Behaviour Research and Therapy,* 1969, **7**, 191–196.

Barlow, D. H., Agras, W. S., Leitenberg, H., & Wincze, J. P. An experimental analysis of the effectiveness of "shaping" in reducing maladaptive avoidance behaviour: An analogue study. *Behaviour Research and Therapy,* 1970, **8**, 165–173.

Boulougouris, J. C., Marks, I. M., & Marset, P. Superiority of flooding (implosion) to desensitization for reducing pathological fear. *Behavior Research and Therapy,* 1971, **9**, 7–16.

Breger, L., & McGaugh, J. L. Critique and reformulation of "learning theory" approaches to psychotherapy and neurosis. *Psychological Bulletin,* 1965, **63**, 338–358.

Clark, D. F. The treatment of monosymptomatic phobia by systematic desensitization. *Behaviour Research and Therapy,* 1963, **1**, 63–68.

Craig, K. D. Physiological arousal as a function of imagined, vicarious, and direct stress experiences. *Journal of Abnormal Psychology,* 1968, **73**, 513–520.

Flanagan, J. Galvanic skin response: Emotion or attention? *Proceedings of the 75th Annual Convention of the American Psychological Association,* 1967, **2**, 7–8.

Folkins, C. H., Lawson, K. D., Opton, E. M., & Lazarus, R. S. Desensitization and the experimental reduction of threat. *Journal of Abnormal Psychology,* 1968, **73**, 100–113.

Gale, E. N., Hyman, E., Ayer, W. Physiological measures during desensitization: A report of two cases. *Journal of Clinical Psychology,* 1970, **26**, 247–250.

Geer, J. H. Fear and autonomic arousal. *Journal of Abnormal Psychology,* 1966, **71**, 253–255.

Geissman, P., & Noel, C. EEG study with frequency analysis and polygraphy of autogenic training. *Proceedings of the 3rd World Congress of Psychiatry,* 1961, **3**, 468–472.

Gelder, M. G., & Mathews, A. M. Forearm blood flow and phobic anxiety. *British Journal of Psychiatry,* 1968, **114**, 1371–1376.

Graham, F. K., & Clifton, R. K. Heart rate as a component of the orienting response. *Psychological Bulletin,* 1966, **65**, 305–320.

Grings, W. W., & Uno, T. Counterconditioning: Fear and relaxation. *Psychophysiology,* 1968, **4**, 479–485.

Grossberg, J. M. The physiological effectiveness of brief training in differential muscle relaxation. (Tech. Rep. No. 9) La Jolla, Calif.: Western Behavioral Sciences Institute, 1965.

Grossberg, J. M., & Wilson, H. K. Physiological changes accompanying the visualization of fearful and neutral situations. *Journal of Personality and Social Psychology,* 1968, **10**, 124–133.

Hoenig, J., & Reed, G. F. The objective assessment of desensitization. *British Journal of Psychiatry,* 1966, **112**, 1279–1283.

Jacobson, E. Electrical measurements concerning muscular contraction (tonus) and the cultivation of relaxation in man. *American Journal of Physiology,* 1934, **107**, 230–248.

Jacobson, E. *Progressive relaxation.* Chicago: University of Chicago Press, 1938.

Jacobson, E. Variation of blood pressure with skeletal muscle tension and relaxation. *Annals of Internal Medicine,* 1939, **12**, 1194–1212.

Jacobson, E. Variation of pulse rate with skeletal muscle tension and relaxation. *Annals of Internal Medicine,* 1940, **13**, 1619–1625.

Jacobson, E. The effects of daily rest without training to relax on muscular tonus. *American Journal of Psychology,* 1942, **55**, 248–254.

Johnson, A. J., & Campos, J. J. The effect of cognitive tasks and verbalization instructions on heart rate and skin conductance. *Psychophysiology,* 1967, **4**, 143–150.

Jus, A., & Jus, K. Some remarks on "passive" concentration and on autogenic shift. In L. Chertok (Ed.), *Psychophysiological mechanisms of hypnosis.* Berlin: Springer-Verlag, 1969.

Kondas, O. Reduction of examination anxiety and "stage fright" by group desensitization and relaxation. *Behaviour Research and Therapy,* 1967, **5**, 273–281.

Lacey, J. I. Psychophysiological approaches to the evaluation of psychotherapeutic process and outcome. In E. Rubinstein & M. Parloff (Eds.), *Research in psychotherapy*. Vol. 1. Washington, D. C.: American Psychological Association, 1959.

Lacey, J. I. Somatic response patterning and stress: Some revisions of activation theory. In M. Appley & R. Trumbull (Eds.), *Psychological stress*. New York: Appleton-Century-Crofts, 1967.

Lader, M. H., Gelder, M. G., & Marks, I. M. Palmar conductance measures as predictors of response to desensitization. *Journal of Psychosomatic Research*, 1967, **11**, 283–290.

Lader, M. H., & Mathews, A. M. A physiological model of phobic anxiety and desensitization. *Behaviour Research and Therapy*, 1968, **6**, 411–421.

Lader, M. H., & Mathews, A. M. Physiological changes during spontaneous panic attacks. *Journal of Psychosomatic Research*, 1970, **14**, 377–382.

Lang, P. J. The mechanics of desensitization and the laboratory study of human fear. In C. Franks (Ed.), *Behaviour therapy: Appraisal and status*. New York: McGraw-Hill, 1969.

Lang, P. J. Stimulus control, response control, and the desensitization of fear. In D. J. Lewis (Ed.), *Learning approaches to therapeutic behaviour*. Chicago: Aldine Press, 1970.

Lang, P. J., & Lazovik, A. D. The experimental desensitization of a phobia. *Journal of Abnormal and Social Psychology*, 1963, **66**, 519–525.

Lang, P. J., Melamed, B. G., & Hart, J. A psychophysiological analysis of fear modification using an automated desensitization procedure. *Journal of Abnormal Psychology*, 1970, **76**, 220–234.

Lazarus, R. S. Emotions and adaptation: Conceptual and empirical relations. *Nebraska Symposium on Motivation*, 1968, **16**, 175–266.

Leitenberg, H., Agras, W. S., & Barlow, D. H. Contribution of selective positive reinforcement and therapeutic instructions to systematic desensitization therapy. *Journal of Abnormal Psychology*, 1969, **74**, 113–118.

Lomont, J. F., & Edwards, J. E. The role of relaxation in systematic desensitization. *Behaviour Research and Therapy*, 1967, **5**, 11–25.

Mathews, A. M., & Gelder, M. G. Psychophysiological investigations of brief relaxation training. *Journal of Psychosomatic Research*, 1969, **13**, 1–12.

McCurdy, H. G. Consciousness and the galvanometer. *Psychological Review*, 1950, **57**, 322–327.

Meyer, V. The treatment of two phobic patients on the basis of learning principles. *Journal of Abnormal and Social Psychology*, 1957, **55**, 261–266.

Miller, N. E. Learnable drives and rewards. In S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley, 1951.

Moore, N. Behaviour therapy in bronchial asthma: A controlled study. *Journal of Psychosomatic Research*, 1965, **9**, 257–276.

Paul, G. L. *Insight vs. desensitization in psychotherapy: An experiment in anxiety reduction.* Stanford: Stanford University Press, 1966.

Paul, G. L. Inhibition of physiological response to stressful imagery by relaxation training and hypnotically suggested relaxation. *Behaviour Research and Therapy*, 1969, **7**, 249–256. (a)

Paul, G. L. Outcome of systematic desensitization. In C. Franks (Ed.), *Behaviour therapy: Appraisal and status*. New York: McGraw-Hill, 1969. (b)

Paul, G. L. Physiological effects of relaxation training and hypnotic suggestion. *Journal of Abnormal Psychology*, 1969, **74**, 425–437. (c)

Rachman, S. Systematic desensitization. *Psychological Bulletin*, 1967, **67**, 93–103.

Rimm, D. C., & Bottrell, J. Four measures of visual imagination. *Behaviour Research and Therapy*, 1969, **7**, 63–69.

Rowland, L. W. The somatic effects of stimuli graded in respect of their exciting character. *Journal of Experimental Psychology*, 1936, **19**, 547–560.

Schultz, J. H., & Luthe, W. *Autogenic training*. New York: Grune and Stratton, 1959.

Seager, C. P., & Brown, B. H. An indicator of tension during reciprocal inhibition. *British Journal of Psychiatry*, 1967, **113**, 1129–1132.

Sokolov, Y. N. *Perception and the conditioned reflex*. New York: Macmillan, 1963.

Valins, S., & Ray, A. A. Effect of cognitive desensitization on avoidance behavior. *Journal of Personality and Social Psychology*, 1967, **7**, 345–350.

Van Egeren, L. F. Psychophysiology of systematic desensitization: Individual differences and the habituation model. *Journal of Behavior Therapy and Experimental Psychiatry*, in press.

Van Egeren, L. D., Feather, B. W., & Hein, P. L. Desensitization of phobias: Some psychophysiological propositions. *Psychophysiology*, in press.

Wilkins, W. Desensitization: Social and cognitive factors underlying the effectiveness of Wolpe's procedure. *Psychological Bulletin*, in press.

Wilson, G. D. An electrodermal technique for the study of phobias. *New Zealand Medical Journal*, 1966, **65**, 696–698.

Wilson, G. D. GSR responses to fear-related stimuli. *Perceptual and Motor Skills*, 1967, **24**, 401–402.

Wolpe, J. *Psychotherapy by reciprocal inhibition*. Stanford: Stanford University Press, 1958.

Wolpe, J., & Flood, J. The effect of relaxation on the galvanic skin response to repeated phobic stimuli in ascending order. *Journal of Behaviour Therapy and Experimental Psychiatry*, 1970, **1**, 195–200.

Zeisset, R. M. Desensitization and relaxation in the modification of psychiatric patients interview behaviour. *Journal of Abnormal Psychology*, 1968, **73**, 18–24.

(Received February 26, 1970)

Psychological Bulletin
1971, Vol. 76, No. 2, 92–104

TEST ANXIETY AND DIRECTION OF ATTENTION [1]

JERI WINE [2]

University of Waterloo

The literature reviewed suggests an attentional interpretation of the adverse effects which test anxiety has on task performance. During task performance the highly test-anxious person divides his attention between self-relevant and task-relevant variables, in contrast to the low-test-anxious person who focuses his attention more fully on the task. This interpretation was supported by literature from diverse areas suggesting that (*a*) highly anxious persons are generally more self-preoccupied than are people low in anxiety; (*b*) the self-focusing tendencies of highly test-anxious persons are activated in testing situations; (*c*) those situational conditions in which the greatest performance differences occur are ones which elicit the self-focusing tendencies of highly test-anxious subjects, and the task-focusing tendencies of low-anxious subjects; (*d*) research examining the relationship between anxiety and task variables suggests that anxiety reduces the range of task cues utilized in performance; (*e*) "worry," an attentionally demanding cognitive activity, is more debilitating of task performance than is autonomic arousal. Treatment and research implications of this attentional interpretation of test anxiety are briefly discussed.

Highly test-anxious persons typically perform more poorly on tests than do low-test-anxious persons, particularly when the tests are administered under stressful, evaluative conditions. The literature reviewed in this article suggests that this performance difference is largely due to a difference in the attentional focuses of high- and low-test-anxious persons during task performance. The low-test-anxious person is focused on task-relevant variables while performing tasks. The highly test-anxious subject is internally focused on self-evaluative, self-deprecatory thinking, and perception of his autonomic responses. Since the difficult tasks on which the test-anxious person does poorly require full attention for adequate performance, he cannot perform adequately while dividing his attention between internal cues and task cues. Marlett and Watson (1968) have stated this proposition rather well:

> The high-test-anxious person spends a part of his task time doing things which are not task oriented. He worries about his performance, worries about how well others might do, ruminates over choices open to him, and is often repetitive in his attempts to solve the task. Any effort to overcome the self-defeating kind of behavior which the high-anxious person exhibits should concentrate on allowing him to perform without the constraints which maximize the negative effects of the avoidance behavior [p. 203].

TEST ANXIETY THEORY AND MEASUREMENT

The ideas presented above are restatements of test anxiety theory as advanced by its originators (Mandler & Sarason, 1952) and later contributors. Test anxiety theory was initially presented by Mandler and Sarason in their article introducing the Test Anxiety Questionnaire (TAQ). In contrast to general anxiety scales such as the Taylor Manifest Anxiety Scale (MAS; Taylor, 1953), the TAQ was constructed as a measure of anxiety proneness in a specific kind of stressful situation—the testing situation. It was expected that items which related specifically to the subject's reactions to testing situations would be more predictive than general anxiety scales of his behavior in these situations.

---

[1] This review is based primarily on a dissertation submitted to the University of Waterloo in partial fulfillment of the requirements for a doctoral degree.

[2] The author wishes to thank the members of her dissertation committee, Marvin Brown, Chairman, Donald Meichenbaum, and Herbert Lefcourt for their helpful suggestions and advice. Particular gratitude is expressed to my husband John J. Wine, for suggesting the basic theoretical approach.

Requests for reprints should be sent to Jeri Wine, Department of Psychology, University of Waterloo, Waterloo, Ontario, Canada.

It was assumed that two kinds of drives are evoked in the testing situation. The first of these are learned task drives, which are reduced by "responses or response sequences which lead to completion of the task [Mandler & Sarason, 1952, p. 166]." The second kind is a learned anxiety drive, which can elicit two classes of responses: those related to task completion, which are anxiety reducing, and those which interfere with task completion. This latter response class is the one which the TAQ was constructed to measure:

> These responses . . . may be manifested as feelings of inadequacy, helplessness, heightened somatic reaction, anticipations of punishment or loss of status and esteem, and implicit attempts at leaving the test situation. It might be said that these responses are self rather than task centered [p. 166].

The task-relevant responses are ones which are specific to the task, and must be learned during task performance. The self-oriented responses are a class of generalized responses which are readily evoked in a task situation, and interfere with the learning of specific task-relevant responses. The TAQ, a 37-item questionnaire in a rating-scale format, is designed to measure these self-oriented responses. The items refer to the kinds of internal responses typically experienced by the subject immediately before and during examinations and tests.

I. Sarason (1958a) constructed a 21-item measure of test anxiety, which he labeled the Test Anxiety Scale (TAS). It is based largely on items taken from Mandler and S. Sarason's TAQ rewritten for a true-false format. The TAQ and the TAS are highly correlated, product-moment correlation = .93 (I. Sarason, Pederson, & Nyman, 1968). I. Sarason has restated the interfering response hypothesis, emphasizing that it is a habit interpretation of anxiety:

> This interpretation, briefly put, states that *Ss* scoring high and low in anxiety differ in the response tendencies activated by personally threatening conditions. Whereas low scoring *Ss* may react to such conditions with increased effort and attention to the task at hand, high scoring *Ss* respond to threat with self-oriented, personalized responses [1960, p. 405].

As Sarason noted, the attentional focuses of high- and low-anxious subjects differ when under threat. The low-anxious subjects turn their attention to the task, while high-anxious subjects attend to their internal self-oriented responses.

Alpert and Haber (1960) constructed a test anxiety questionnaire, the Achievement Anxiety Test (AAT), which yields two measures. One of these is a 10-item debilitating anxiety scale, a construct similar to that measured by the TAQ and TAS. Total scores on the scale correlated .64 with total TAQ scores. The other measure is labeled facilitating anxiety. Mandler and S. Sarason (1952) had theorized that test anxiety is debilitating of performance only among subjects who had learned a habitual class of interfering responses to the test anxiety. Among persons without this class of interfering responses, test anxiety elicits task-relevant responses leading to task completion. Persons scoring low on the TAQ were implicitly assumed by Mandler and Sarason to be high in the facilitating kind of anxiety. Alpert and Haber's AAT measures separately the two constructs of anxiety and their associated responses. The nine items on the facilitating anxiety scale are based on "a prototype of the item—'Anxiety helps me to do better during examinations and tests [p. 213].'" The facilitating and debilitating scales correlated −.48 with each other (Alpert & Haber, 1960).

Throughout this review, the terms "high test anxious" and "low test anxious" refer to persons who score at the high and low extremes of measures of debilitating test anxiety, such as the TAQ, TAS, or debilitating anxiety scale of the AAT.

SELF-FOCUSING TENDENCIES OF HIGH-ANXIOUS SUBJECTS: EMPIRICAL EVIDENCE

In a review of the literature on paper-and-pencil anxiety scales, I. Sarason (1960) cited a number of studies that provide evidence that high-anxious subjects are "more self-deprecatory, more self-preoccupied, and generally less content with themselves than subjects lower in the distribution of anxiety scales [p. 404]." This conclusion was primarily based on the results of studies which reported relationships between scores on anxiety scales and other paper-and-pencil personality measures. In general, these investigations indicated that persons who score high on anxiety scales describe themselves in negative, self-devaluing terms on other pencil-and-paper measures as well.

Two of the studies reviewed in this connection by I. Sarason reported relationships between test anxiety and responses of subjects in experimental situations. In one of these (Doris & S. Sarason, 1955) subjects differing in TAQ scores were arbitrarily failed on a number of tasks. Following the tasks, the subjects were required to rank order "blame statements," which included "self-blame" and "other-than-self" blame items. Highly anxious subjects blamed themselves for their failures significantly more than did low-anxious subjects. In the other study, Trapp and Kausler (1958) compared Wechsler-Bellevue digit-symbol performance and levels of aspiration of high- and low-test-anxious subjects. Through the actual performance of the high- and low-TAQ groups did not differ, the levels of aspiration of high-anxious subjects became progressively lower over the four trials. On the last two trials, the high-test-anxious subjects had significantly lower level-of-aspiration scores than the low-anxious subjects. Though they did not objectively perform more poorly than the low-TAQ subjects, the high-anxious subjects progressively became more pessimistic about their performance over the four trials.

A more recent study by Meunier and Rule (1967) reported results leading to similar conclusions regarding the test-anxious person's tendency to devalue his own performance. The effects of positive, negative, or no feedback on the subjects' confidence in their judgments of the length of lines were investigated. On no-feedback trials, highly test-anxious subjects rated their confidence level as low and comparable to their confidence on trials with negative feedback. In contrast, low-test-anxious subjects expressed high confidence in their judgment on no-feedback trials, and this level of confidence corresponded to the level expressed on positive feedback trials.

Since publishing his review of anxiety scale studies, I. Sarason has completed several studies which further indicate the self-deprecatory, self-ruminative tendencies of highly test-anxious subjects (Sarason & Ganzer, 1962, 1963; Sarason & Koenig, 1965). In each of these studies, an unstructured verbal conditioning paradigm was used with college students who scored at the extremes of the TAS. The subjects were required to describe themselves orally for approximately one-half hour. Nonreinforcement, reinforcement of negative self-references, and reinforcement of positive self-references have been compared. Some of the conclusions from these experiments are: (*a*) Regardless of experimental condition, highly test-anxious subjects generally describe themselves in more negative terms than do low-test-anxious subjects. (*b*) High-test-anxious subjects are extremely responsive to reinforcement when the response class being reinforced is negative self-references. (*c*) However, when the response class being reinforced is positive self-references, high-anxious subjects do not condition. That is, they do not produce more positive self-references as a result of being verbally reinforced for them.

The conclusion that high-anxious persons are, indeed, generally self-deprecatory and self-preoccupied seems inescapable in light of the research evidence. But whether these self-focusing tendencies are specifically activated in testing situations is a question which has not been answered in the studies reviewed to this point. Four recent studies (Ganzer, 1968; Mandler & Watson, 1966; Marlett & Watson, 1968; Neale & Katahn, 1968) provide evidence bearing directly on this proposition.

In the earliest of these studies, Mandler and Watson (1966) administered a series of digit-symbol tasks to extreme low- and high-TAQ groups. A posttask questionnaire included the question, "How often during the testing did you find yourself thinking how well, or how badly, you seemed to be doing? [p. 276]." On a 10-point rating scale, high-test-anxious subjects indicated markedly greater occurrence of such thoughts than did the low-test-anxious group. In a partial replication of this study, Neale and Katahn (1968) reported identical results on this questionnaire item.

Marlett and Watson (1968) reported a similar result. Ninth-grade males in high- and low-extreme groups on the children's form of the Alpert-Haber Achievement Anxiety Test

(Stanford, Dember, & Stanford, 1963) were arbitrarily failed for 12 trials on a button-pressing task. A posttask questionnaire included the question, "How often did you think about how well or badly you were doing?" The ends of the rating scale were labeled "I didn't think about it," and "I thought about it so much I couldn't concentrate [pp. 202–203]." The scores of high-test-anxious subjects indicated that they were significantly more disturbed by such self-focused thoughts.

Even more direct evidence for the activation of self-oriented tendencies among high-anxious subjects in test situations is provided in the study reported by Ganzer (1968). The effects of audience presence and test anxiety (TAS) on the serial verbal learning of female subjects were investigated. Tabulations were made of the frequency and content of all of the subjects' task-irrelevant comments while they were working on the task. "High TAS scorers, especially in the Observed condition, emitted more than any other group. Content analysis revealed that the comments were mostly of a self-evaluative or apologetic nature [Ganzer, 1968, p. 194]."

In summary, there is abundant evidence that the test-anxious person is more self-preoccupied and self-deprecatory than his low-test-anxious counterpart. It is also clear, in view of the research reviewed, that these self-focused tendencies are activated by the pressures of the testing situation.

SITUATIONAL CONDITIONS AND THE TASK PERFORMANCE OF HIGH- AND LOW-TEST-ANXIOUS SUBJECTS

The two initial studies published by the authors who originated test anxiety theory (Mandler & S. Sarason, 1952; S. Sarason, Mandler, & Craighill, 1952) were prototypes of the many subsequent investigations of situational stress variables and test anxiety. The general predictions of test anxiety theory with regard to situational stress variables are:

> When a stimulus situation contains elements which specifically arouse test or achievement anxiety, this increase in anxiety drive will lead to poorer performance in individuals who have task-irrelevant responses in their response repertory. For individuals without such response tendencies, these stimulus elements will raise their general drive level and result in improved performance. (Such elements would be any reference that the individual is being judged, a statement of expected performance, etc.) [Sarason et al., 1952, p. 561].

Mandler and Sarason (1952) examined the effects of feedback and no-feedback on the performance of subjects scoring at extremes on the TAQ. As predicted, the two feedback conditions, both success and failure, facilitated the performance of low-TAQ subjects, and interfered with the performance of high-TAQ subjects. In the second study (Sarason et al., 1952) similar predictions were made concerning an instructional manipulation. In the stressful "ego-involving" instructional condition, subjects were told that they should easily complete the test within the time limit, when it was, in fact, made impossible for them to do so. This "expected-to-finish" condition can be described as an advance failure manipulation. The nonstressful instructions, "not-expected-to-finish," informed the subject that no one could expect to finish within the time limit. In general, the high-TAQ subjects did better following the "not-expected-to-finish" than the "expected-to-finish" instructions, while the reverse was true for the low-anxious subjects.

The many subsequent investigations of stressful situational conditions and test anxiety have generally provided data consistent with the predictions of test anxiety theory (Cox, 1966, 1968; Ganzer, 1968; Harleston, 1962; Paul & Eriksen, 1964; Russell & I. Sarason, 1965; I. Sarason, 1958b, 1959, 1961; I. Sarason & Harmatz, 1965; I. Sarason & Minard, 1962; I. Sarason & Palola, 1960). Most of these studies have investigated the effects of feedback or instructional manipulations, though three were concerned with the effects of audience presence on task performance (Cox, 1966, 1968; Ganzer, 1968). Considering audience presence as a source of threatening evaluation, the results of both studies support the predictions of test anxiety theory. Audience presence debilitated the task performance of high-test-anxious subjects, and facilitated the performance of low-test-anxious subjects.

Instructional variations, other than those discussed above, have most frequently involved comparisons of (*a*) "ego-involving," "highly motivating" instructions, which inform the subject that his task performance will reflect his ability, (*b*) minimal task instructions, and (*c*) instructions which assure the subject that his task performance will not be used to evaluate him. Frequently, the subject is told that his performance will be anonymous. He may be instructed that the experimenter is interested only in properties of the task, or that his performance is important only so that the experimenter may complete the study.

The results of studies varying instructional conditions have generally reported an interaction between level of test anxiety and evaluation emphasis: (*a*) Highly test-anxious subjects perform more poorly following highly evaluative "ego-involving" instructions than nonevaluative "anonymous" instructions while the reverse is true for low-test-anxious subjects. (*b*) Following highly evaluative instructions, low-test-anxious subjects perform better than high-test-anxious subjects; following nonevaluative instructions, high-test-anxious subjects perform better than low. (*c*) Following minimal task instructions, high- and low-test-anxious-subjects perform at about equivalent levels, intermediate between their performances in the highly evaluative and nonevaluative conditions.

All of the situational conditions which differentially affect the performance of high- and low-anxious subjects vary on evaluative dimension—the degree to which the subject believes that he is being or will be evaluated on the basis of his task performance. Conditions at the extremes of the dimension affect the performance of high- and low-test-anxious persons in opposing ways. It might also be noted that situational conditions varying on this evaluative dimension have opposing effects on the direction of attention of high- and low-test-anxious persons. Highly evaluative conditions cause the highly test-anxious person to direct his attention internally rather than to the task. The same conditions cause the low-test-anxious person to increase his "attention to the task at hand [I. Sarason, 1960, p. 405]." Nonevaluative situational conditions do not elicit the highly test-anxious subject's self-directed interfering responses, thus it is possible for him to direct his full attention to the task. Conversely, nonevaluative testing conditions presumably do not excite the nonanxious subject's motivation and interest. Though it cannot be said that he therefore turns his attention inward, it is probable that his low interest in the task is accompanied by less than full attention to it.

I. Sarason (1958b) has investigated another instructional variation, which he labeled a "reassurance" condition. These instructions, like the nonevaluative ones discussed above, were designed to improve the performance of high-test-anxious subjects rather than interfere with it. In spite of the "reassurance" label, these instructions can most accurately be described as neutral with respect to the evaluative dimension, since there was no reference in them to evaluation of the subject's performance. Their nature might more accurately be described as information-giving and attention-directing. The subject's task was to learn a serial list of nonsense syllables. Following the standard serial learning task instructions, the subject was warned that many people become tense when learning this kind of task and that he should expect his progress to be slow. He was further told that he should concentrate on the list rather than worrying about how he was doing. When this set of instructions was compared to standard or minimal task instructions, high-test-anxious subjects performed better following the reassurance instructions. The reassurance condition was detrimental to the performance of the low-test-anxious subjects, in comparison to the standard task instructions. Sarason's study suggests that it may not be necessary to resort to removing the evaluative emphasis from the highly anxious subject's performance in order to improve it. The results indicate that it may be sufficient to substitute task-relevant responses for the self-orienting ones by means of rather straightforward attention-directing instructions.

ANXIETY AND RANGE OF CUE UTILIZATION

It was suggested in the preceding section that evaluative testing conditions have opposite effects on the attentional focuses of high- and low-test-anxious persons. When being evaluated, the high-test-anxious person turns his attention inward while the low-test-anxious person focuses more fully on the task. The implication is that the high-test-anxious person attends to fewer task cues than does the low-test-anxious person. This attentional interpretation of the effects of test anxiety on task performance is consistent with an empirical generalization advanced by Easterbrook (1959) concerning the relationship between arousal level and task variables. He reviewed a large body of research which indicated that emotional arousal consistently narrows the range of cue utilization in task performance.

The following are examples of the research Easterbrook cited to support this generalization: (*a*) Studies of the effects of experimentally manipulated drive level on response to a focal task, and simultaneous response to peripheral, only occasionally relevant stimuli. Arousal reduces responsivity to the peripheral stimuli, while maintaining or improving central performance (Bahrick, Fitts, & Rankin, 1952; Bursill, 1958; Davis, 1948; Easterbrook, 1953). (*b*) Studies of the effects of emotional arousal on measures of incidental learning. Again arousal reduces the amount of incidental learning, while having no effect on or improving intentional learning (Aborn, 1953; Bahrick, 1954; Bruner, Matter, & Papanek, 1955; Johnson, 1952; Kohn, 1954; Silverman, 1954; Silverman & Blitz, 1956). These are only a few of the studies cited by Easterbrook in support of his formulation.

Easterbrook's definition of emotional arousal is very broad:

> the innate response to a state of biological deprivation or noxious stimulation, which underlies or occurs simultaneously with overt action and affects its strength and course. This emotional arousal is greater in neurotic than in normal subjects, greater than usual in subjects under stress or threat or in frustration, and in general greater in animals that have been "motivated" by any of the usual deprivations, noxious stimulations, or other incentives than it is in unmotivated or resting animals of the same species [p. 184].

Anxiety is one aspect of this emotional arousal dimension, and several of the studies reviewed by Easterbrook indicate that narrowing the range of cue utilization accompanies anxiety arousal.

Since Easterbrook's review, several studies have reported the effects of experimentally aroused anxiety on range of cut utilization in task performance (Agnew & Agnew, 1963; Bruning, Capage, Kozuh, Young, & Young, 1968; Tecce & Happ, 1964; Wachtel, 1966, 1968). The results of these investigations have been consistent with Easterbrook's formulation. Stressful situational conditions tended to reduce the range of cue utilization, in comparison to nonstressful conditions. No measures of individual differences were used in these studies.

Zaffy and Bruning (1966), West, Lee, and Anderson (1969) and Wachtel (1966, 1968) reported relationships between scale-measured anxiety level and range of cue utilization. Zaffy and Bruning selected subjects from extreme scorers on the Taylor MAS. Rather than varying range of cues, they made the available cues either relevant or irrelevant to task performance. The subjects' task was to learn a correct position for each of 19 presentations of 5 zeros, presented by a memory drum. In the relevant cue condition the zeros were numbered in order with subscripts from 1 to 5; the same subscripts were used in the irrelevant cue condition but in random order. These conditions were compared to a no-cue control condition. The performance of the low-MAS subjects was affected more by the presence of the cues, whether relevant or irrelevant, than was the performance of the high-MAS subjects. Within the low-MAS anxiety level, there were differences significant beyond the .001 level in all three of the comparisons of the three conditions, with superior performance in the relevant cue condition, intermediate performance in the no-cue condition, and poorest performance in the irrelevant cue condition. The performance levels of the high MAS subjects followed the same order, but there was a significant difference only between the relevant and irrelevant cue conditions ($p < .025$).

More pertinent to the purposes of this review are the studies reported by West et al. (1969) and Wachtel (1966, 1968), since they selected subjects on the basis of their scores on measures of test anxiety. West et al. used sixth and eighth graders who were high and low scorers on the Test Anxiety Scale for Children (TASC; S. Sarason, Davidson, Lighthall, Waite, & Ruebush, 1960). The task was made up of arithmetic problems. There were two conditions determined by the nature of the information given in the problems. In one condition only information essential to solution of the problems was given; in the other condition the same relevant information was provided, but additional irrelevant information was given in each problem. The interaction between anxiety level and condition was significant. Both high- and low-TASC groups performed more poorly in the irrelevant information condition. No tests of significance for individual pairs of means were reported; the means for the groups indicate that the performance level of the low-TASC group was affected more by the addition of irrelevant cues than was the performance level of the high-TASC group: low TASC, relevant information only, $\bar{X} = 33.4$, relevant plus irrelevant, $\bar{X} = 23.8$; high TASC, relevant only, $\bar{X} = 27.3$, relevant plus irrelevant, $\bar{X} = 24.4$. The low-test-anxious children were more sensitive to the addition of irrelevant cues than were the high-test-anxious children.

Wachtel's study (1966, 1968) reported relationships between test anxiety level and range of cue utilization in task performance on a central continuous tracking task and reaction times to two occasional peripheral lights. A control condition with minimal task instructions was compared with three experimental conditions: (*a*) Subjects told that unavoidable electric shock would be administered during task performance; (*b*) an avoidable shock condition in which subjects were informed that they could avoid shock if their combined performance on the central and peripheral tasks remained high; (*c*) a personality diagnostic condition in which the task was presented as a means of evaluating the subject's personality; the subject was told that the peripheral lights would go on only if his performnce fell below an acceptable level.

Though Wachtel's study was primarily focused on the effects of experimentally manipulated stress on performance, he reported interesting relationships between subjects' TAQ scores and task performance. The "avoidable shock" and "personality diagnosis" conditions were the only ones in which test anxiety was appropriately engaged through threatening the subject with evaluation of his performance. In both of these conditions there was evidence of narrowing attention to task cues as a result of test anxiety. In the avoidable shock condition there was a nearly significant r of $-.35$ ($p < .10$) between speed scores (reciprocals of reaction times to the occasional peripheral lights) and subjects' TAQ scores, and a significant positive correlation between TAQ scores and scores on the central tracking task ($r = .58$, $p < .01$). In the personality diagnosis condition the correlation between speed scores and TAQ scores was $-.44$ ($p < .05$), though the correlation between TAQ scores and tracking scores in this condition was only .02. When all subjects in these two conditions were combined, the correlation between TAQ scores and speed scores was $-.40$ ($n = 30$, $p < .01$). Thus in these two conditions test anxiety reduced responsiveness to the peripheral lights. In the other two conditions, in which test anxiety was not appropriately aroused, there were no significant correlations between TAQ scores and task performance.

A proposition offered by Wachtel to explain the effects of experimentally manipulated anxiety and test anxiety in this study combines Easterbrook's attentional formulation and the interfering response hypothesis of the test anxiety theorists: "when an individual is anxious, attention is diverted inward to perception of his anxiety and therefore less attention is available for external stimuli [Wachtel, 1966, p. 2149]."

WORRY AND EMOTIONALITY: COMPONENTS OF TEST ANXIETY

Liebert and Morris (1967) have presented an analysis of test anxiety which is quite consistent with the one proposed in this paper. They suggest that test anxiety is composed of two major aspects: worry and emotionality. The worry (W) component is described as cognitive concern over performance, and emotionality (E) is the autonomic arousal aspect of anxiety. Liebert and Morris devised a brief situational measure of W and E based on TAQ items. The results of a series of studies (Doctor & Altman, 1969; Liebert & Morris, 1967; Morris & Liebert, 1969, 1970; Spiegler, Morris, & Liebert, 1968) suggest that scores on the two components vary predictably with temporal relationships to classroom examinations and with performance expectancies. The scale has been administered several days before an examination, immediately before, and immediately after. Scores on W tend to be fairly constant across time; while E scores reach a peak immediately before an examination, falling off rapidly immediately after the examination. W scores are significantly and negatively correlated with subjects' preexamination ratings of performance expectancy, while E scores are not related to performance expectancy.

Three studies in this series have yielded results on the relationship between worry and emotionality and task performance (Doctor & Altman, 1969; Morris & Liebert, 1969, 1970). Morris and Liebert (1969) devised another paper-and-pencil measure of the two anxiety components, this one composed of items from the Taylor MAS. The performance measure was a total of five Wechsler Adult Intelligence Scale (WAIS) subtests. A 2 × 2 factorial design was used: half of the subjects completed only the easy items in each subtest, half the difficult items; half of the subjects were aware of being timed, half were not. There were significant interactions between worry scores and both task difficulty and timing in their effects on total WAIS scores. There were no significant results involving emotionality and task performance. The authors suggest that "it is worry, not 'anxiety,' which affects performance on intellectual-cognitive tasks and which interacts with the relevant variables of the test situation [pp. 243–244]."

Doctor and Altman (1969) examined the relationships between W and E scores on Liebert and Morris' (1967) scale and final examination grades in a psychology course. W scores were found to be significantly and negatively related to exam performance. E scores were negatively related to performance only among subjects below the group median W score. When W was high, the addition of E had little or no effect on performance.

Morris and Liebert (1970) reported similar results in two studies, one involving college undergraduates, the other, high school students. Partial correlations were computed between worry and emotionality scores on the Liebert-Morris (1967) scale, which was administered immediately before a course final examination and examination grades. In both studies, partial correlations between worry scores and final examination grades were negative and significant. Partial correlations between emotionality scores and grades were nonsignificant.

The worry component seems closely related to an attentional interpretation of test anxiety, which proposes that the adverse effects of test anxiety are due to attention being divided between self and the task. Worry has been described as "preoccupation with performance [Doctor & Altman, 1969, p. 564]," "*cognitive concern* about the consequences of failing, the ability of others relative to one's own, etc. [Liebert & Morris, 1967, p. 975]," "cognitive or intellectual concern about one's own performance [Spiegler et al., 1968, p. 451]." These descriptions all refer to attentionally demanding cognitive activity going on during task performance. This is not to say that autonomic activity is not demanding of attention as well, but it is less likely to require attention except "at high levels where physiological reactivity might be distractive and annoying [Doctor & Altman, 1969, p. 364]."

CONCLUSIONS AND IMPLICATIONS

The literature reviewed supports an attentional interpretation of the debilitating effects of test anxiety on task performance. The highly test-anxious person responds to evaluative testing conditions with ruminative, self-evaluative worry and, thus, cannot direct adequate attention to task-relevant variables. An attentional analysis is closely related to existing theories of test anxiety. Mandler and Sarason's (1952) original theoretical position assumed that anxiety interferes with the performance only of subjects who have a habitual class of self-oriented interfering responses in their response repertoires. For subjects without this class of interfering responses, anxiety may facilitate task-relevant responses. Alpert and Haber's Achievement Anxiety Test was designed to provide independent measures of these two classes of response to test anxiety. Liebert and Morris'

(1967) analysis of test anxiety into the two components of worry and emotionality is a logical extension of test anxiety theory. Emotional arousal appears to bear no consistent relationship to performance on intellectual or cognitive tasks, while worry is consistently and negatively related to performance. An attentional interpretation states simply that the reason "worry" debilitates task performance is that it is attentionally demanding and distracts attention from the task.

Before moving into specific research and treatment applications, a few implications of the attentional analysis of test anxiety are stated below:

1. An attentional approach is explicitly concerned with how the subject uses his task time—his cognitive activity, what he is thinking about and attending to.

2. This approach implies little interest in autonomic arousal per se. In this context, degree of arousal is irrelevant unless the subject is attending to his arousal. Of course, as noted earlier, when arousal becomes quite extreme, it is attentionally demanding.

3. Finally, this analysis implies that the test-anxious person's performance may be improved by directing his attention to task-relevant variables, and away from self-evaluative rumination.

Experimental Applications

Though test anxiety theorists have not viewed test anxiety as a unitary dimension of emotional arousal, the experimental studies in this area have typically been concerned with manipulations of the evaluative situational conditions which evoke test anxiety. With two notable exceptions (Sarason et al., 1968; Sieber, 1969) studies have not investigated conditions designed to alter the interfering responses elicited by the anxiety. Research manipulating evaluative conditions has been necessary in order to establish test anxiety as a theoretical construct and to validate measures of test anxiety. These goals have been realized; it seems reasonable to suggest that an appropriate next stage for research in this area is one in which the evaluative dimension is held constant, and in which a search is begun for variables that will alter the test-anxious subjects' interfering responses. An attentional analysis of test anxiety can provide some direction to such a search.

There is a large body of literature in selective attention, which has recently been reviewed by Egeth (1967). The basic premise of the studies in this research area is that subjects' attention can be differentially directed to specific stimuli or stimulus attributes and away from others. The nature of the mechanisms involved, the completeness of the selectivity, and the level at which the selection takes place are matters of some dispute (e.g., Broadbent, 1958; Deutsch & Deutsch, 1963; Treisman, 1964). It is, however, empirically established that persons can be instructed to be selectively more attentive to specific stimulus attributes or dimensions and to be less attentive to others. Moreover, with repeated training under attention-directing instructions, subjects become more and more skilled in attending to the "relevant" stimuli and ignoring the "irrelevant" stimuli.

The selective attention literature has not been concerned with individual differences. There is no evidence, therefore, on how test-anxious persons might respond to attention-directing instructions. There is a good deal of evidence, however, which indicates that highly test-anxious persons are very responsive to social cues of all sorts, including verbal reinforcement (I. Sarason & Ganzer, 1962, 1963) modeling cues (I. Sarason et al., 1968), persuasion (Janis, 1955), and conformity pressures (Meunier & Rule, 1967), as well as evaluative instructional manipulations. It seems a plausible hypothesis that test-anxious subjects should be susceptible to attentional instructions directing them to selectively attend to task-relevant variables and to ignore self-relevant variables.

Instructional manipulations are the most obvious means of directing subject's attention to task-relevant variables, but they are not the only available means. For example, modeling has been demonstrated as a very effective means of transmitting new behavior (e.g., Bandura, 1969; Bandura & Walters, 1963). I. Sarason et al (1968) and I. Sarason (1968) have reported studies suggesting that the verbal learning of highly test-anxious subjects is facilitated by prior exposure to businesslike live models learning similar serial lists. The use of models who transmit task-attending, nonworrying cues may be an effective technique for other kinds of learning situations.

Task cues themselves may provide another means of manipulating the attention of test-anxious subjects. Sieber (1969) has reported a study in which high- and low-test-anxious children worked on a multistage task. In one condition, they were provided with visual displays of all of the earlier stages of the task; in another condition, there were no displays. The highly test-anxious children performed better in the visual display than the no-display condition, while the displays had no differential effect on the performance of the low-test-anxious children. Seiber described the displays as "memory supports." They might also be described as means of sustaining attention to task-relevant variables. It should be possible to develop such "memory-supporting" or "attention-sustaining" techniques for use with other tasks and age levels.

To this point, the research possibilities discussed have been concerned only with means of manipulating the attention of highly test-anxious persons. These suggestions do not exhaust the research possibilities implied by an attentional approach. The approach also suggests analysis of other areas, for instance, research designed to examine in detail the manner in which high- and low-test-anxious persons deploy their attention during task performance. Such research would inevitably lead to concern with the properties of specific tasks and the attentional demands they make on subjects.

Treatment Applications

The purpose of this section is not to review in any detail the treatment literature in test anxiety, but rather to suggest possible treatment techniques which might be derived from an attentional approach. The test anxiety treatment literature is described briefly, however, in order to provide some contrast with an attentional treatment approach.

All of the published accounts of treatment of test-anxious subjects have involved variations of systematic desensitization techniques (Cohen, 1969; Donner & Guerney, 1969; Emery & Krumboltz, 1967; Garlington & Cotler, 1968; Ihli & Garlington, 1969; Johnson & Sechrest, 1968; Katahn, Strenger, & Cherry, 1966; Kondaš, 1967; Laxer, Quarter, Kooman, & Walker, 1969; Snider & Oetting, 1966; Suinn, 1968). This literature has contributed a graet deal to analysis of the relative efficacy of various components of the systematic desensitization procedures, for instance, standardized versus individualized anxiety hierarchies, partial versus full hierarchies, relaxation only versus full desensitization, etc. The research indicates in general that systematic desensitization has beneficial effects on self-reported measures of test anxiety, and often, on scholastic performance. Therefore, it is not with any intent to minimize the value of this treatment literature that it is noted that these studies have evolved from interest in specific treatment techniques, rather than from an analysis of the nature and effects of test anxiety. Test anxiety has been considered to be a severe problem, and thus its treatment is considered a stringent test of systematic desensitization procedures. It has implicitly been assumed that test anxiety differs only in degree from the specific anxieties and phobias (e.g., rat, spider, snake) dealt with in much of the behavior modification literature. It should also be noted that by training test-anxious subjects to relax in the presence of progressively more stressful stimuli, a systematic desensitization treatment approach assumes that the emotional arousal component of test anxiety is its defining characteristic.

The attentional formulation suggests quite a different approach to treatment of test-anxious persons. First of all, the specific concern with how highly test-anxious subjects use their task time suggests treatment in which subjects are given intensive practice in dealing with tests. Second, this task practice should be accompanied by instructions to focus fully on the tasks and to inhibit self-relevant thinking. In a pilot study, the author (Wine, 1970) has compared the relative efficacy of attentional training of the kind described, attentional training plus relaxation training, and a "self-explorational" therapy approach with test-anxious undergraduate students. Each treatment condition was conducted over six hourly sessions. Test-anxious subjects in the attentional training condition changed positively and significantly on self-report measures of test-related anxiety, as well as on two standardized performance measures. The performance measures were Forms 1 and 2 of the Wonderlic Personnel Tests (Wonderlic, 1959) and matched forms of a digit-symbol test.[3] The attentional training plus relaxation also produced positive changes, but added nothing beyond attentional training alone. Subjects in the self-explorational condition did not change significantly on any measures.

A larger, more adequate treatment study is being planned for the near future. Though the results of this pilot study are no more than suggestive, one of the suggestions is that attentional training not only improves task performance but lowers reported test anxiety level as well. In conclusion, perhaps it is possible to define test anxiety attentionally. If a person is not attending to his test anxiety, it, in effect, does not exist. "My experience is what I agree to attend to [William James, 1890, p. 403]."

---

[3] The digit-symbol tests were selected from a battery of eight, constructed and standardized by Marvin Brown, University of Waterloo, 1968.

REFERENCES

Aborn, M. The influence of experimentally induced failure on the retention of material acquired through set and incidental learning. *Journal of Experimental Psychology,* 1953, **45**, 225–231.

Agnew, N., & Agnew, M. Drive level effects on tasks of narrow and broad attention. *Quarterly Journal of Experimental Psychology,* 1963, **15**, 58–62.

Alpert, R., & Haber, R. N. Anxiety in academic achievement situations. *Journal of Abnormal and Social Psychology,* 1960, **61**, 207–215.

Bahrick, H. P. Incidental learning under two incentive conditions. *Journal of Experimental Psychology,* 1954, **47**, 170–172.

Bahrick, H. P., Fitts, P. M., & Rankin, R. E. Effects of incentives upon reactions to peripheral stimuli. *Journal of Experimental Psychology,* 1952, **44**, 400–406.

Bandura, A. *Principles of behavior modification.* New York: Holt, Rinehart & Winston, 1969.

Bandura, A., & Walters, R. H. *Social learning and personality development.* New York: Holt, Rinehart & Winston, 1963.

Broadbent, D. E. *Perception and communication.* New York: Pergamon Press, 1958.

Bruner, J. S., Matter, J., & Papanek, M. L. Breadth of learning as a function of drive level and mechanization. *Psychological Review,* 1955, **62**, 1–10.

Bruning, J. L., Capage, J. E., Kozuh, G. F., Young, P. F., & Young, W. E. Socially induced drive and range of cue utilization. *Journal of Personality and Social Psychology,* 1968, **9**, 242–244.

Bursill, A. E. The restriction of peripheral vision during exposure to hot and humid conditions. *Quarterly Journal of Experimental Psychology,* 1958, **10**, 113–129.

Cohen, R. The effects of group interaction and progressive hierarchy presentation on desensitization of test anxiety. *Behaviour Research and Therapy,* 1969, **7**, 15–26.

Cox, F. N. Some effects of test anxiety and presence or absence of other persons on boys' performance on a repetitive motor task. *Journal of Experimental Child Psychology,* 1966, **3**, 100–112..

Cox, F. N. Some relationships between test anxiety, presence or absence of male persons, and boys' performance on a repetitive motor task. *Journal of Experimental Child Psychology,* 1968, **6**, 1–12.

Davis, *D. R. Pilot error.* (Air publication 3139A) London: Her Majesty's Stationery Office, 1948.

Deutsch, J. A., & Deutsch, D. Attention: Some theoretical considerations. *Psychological Review,* 1963, **70**, 80–90.

Doctor, R. M., & Altman, F. Worry and emotionality as components of test anxiety: Replication and further data. *Psychological Reports,* 1969, **24**, 563–568.

Donner, L., & Guerney, B. G., Jr. Automated group desensitization for test anxiety. *Behaviour Research and Therapy,* 1969, **7**, 1–13.

Doris, J., & Sarason, S. B. Test anxiety and blame assignment in a failure situation. *Journal of Abnormal and Social Psychology,* 1955, **50**, 335–338.

Easterbrook, J. A. The measurement of perseverance. Unpublished master's thesis, Queen's University Library, Kingston, Ontario, 1953.

Easterbrook, J. A. The effect of emotion on cue utilization and the organization of behavior. *Psychological Review,* 1959, **66**, 83–201.

Egeth, H. Selective attention. *Psychological Bulletin,* 1967, **1**, 41–57.

Emery, J. R., & Krumboltz, J. D. Standard versus individualized hierarchies in desensitization to reduce test anxiety. *Journal of Counseling Psychology,* 1967, **14**, 204–209.

Ganzer, V. J. Effects of audience presence and test anxiety on learning and retention in a serial learning situation. *Journal of Personality and Social Psychology,* 1968, **8**, 194–199.

Garlington, W. K., & Cotler, S. B. Systematic desensitization of test anxiety. *Behaviour Research and Therapy,* 1968, **6**, 247–256.

Harleston, B. W. Test anxiety and performance in problem-solving situations. *Journal of Personality,* 1962, **30**, 557–573.

Ihli, K. L., & Garlington, W. K. A comparison of group versus individual desensitization of test anxiety. *Behaviour Research and Therapy,* 1969, **7**, 207–209.

James, W. *Principles of psychology.* Vol. 1. New York: Henry Holt, 1890.

Janis, I. L. Anxiety indices related to susceptibility to persuasion. *Journal of Abnormal and Social Psychology*, 1955, **51**, 663–667.

Johnson, E. E. The role of motivational strength in latent learning. *Journal of Comparative and Physiological Psychology*, 1952, **45**, 526–530.

Johnson, S. M., & Sechrest, L. Comparison of desensitization and progressive relaxation in treating test anxiety. *Journal of Consulting and Clinical Psychology*, 1968, **32**, 280–286.

Katahn, M., Strenger, S., & Cherry, N. Group counseling and behavior therapy with test-anxious college students. *Journal of Consulting Psychology*, 1966, **30**, 544–549.

Kohn, H. Effects of variations of intensity of experimentally induced stress situations upon certain aspects of perception and performance. *Journal of Genetic Psychology*, 1954, **85**, 289–304.

Kondaš, O. Reduction of examination anxiety and 'stage-fright' by group desensitization and relaxation. *Behaviour Research and Therapy*, 1967, **5**, 275–281.

Laxer, R. M., Quarter, J., Kooman, A., & Walker, K. Systematic desensitization and relaxation of high-test-anxious secondary school students. *Journal of Counseling Psychology*, 1969, **5**, 446–451.

Liebert, R. M., & Morris, L. W. Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports*, 1967, **20**, 975–978.

Mandler, G., & Sarason, S. B. A study of anxiety and learning. *Journal of Abnormal and Social Psychology*, 1952, **47**, 166–173.

Mandler, G., & Watson, D. L. Anxiety and the interruption of behavior. In C. D. Spielberger (Ed.), *Anxiety and behavior*. New York: Academic Press, 1966.

Marlett, N. J., & Watson, D. Test anxiety and immediate or delayed feedback in a test-like avoidance task. *Journal of Personality and Social Psychology*, 1968, **8**, 200–203.

Meunier, C., & Rule, B. G. Anxiety, confidence and conformity. *Journal of Personality*, 1967, **35**, 498–504.

Morris, L. W., & Liebert, R. M. Effects of anxiety on timed and untimed intelligence tests. *Journal of Consulting and Clinical Psychology*, 1969, **33**, 240–244.

Morris, L. W., & Liebert, R. M. Relationship of cognitive and emotional components of test anxiety to physiological arousal and academic performance. *Journal of Consulting and Clinical Psychology*, 1970, **35**, 332–337.

Neale, J. M., & Katahn, M. Anxiety choice and stimulus uncertainty. *Journal of Personality*, 1968, **36**, 235–245.

Paul, G. L., & Eriksen, C. W. Effects of test anxiety on "real-life" examinations. *Journal of Personality*, 1964, **32**, 480–494.

Russell, D. G., & Sarason, I. G. Test anxiety, sex, and experimental conditions in relation to anagram solution. *Journal of Personality and Social Psychology*, 1965, **1**, 493–496.

Sarason, I. G. Interrelationships among individual difference variables, behavior in psychotherapy and verbal conditioning. *Journal of Abnormal and Social Psychology*, 1958, **56**, 339–344. (a)

Sarason, I. G. The effects of anxiety, reassurance, and meaningfulness of material to be learned in verbal learning. *Journal of Experimental Psychology*, 1958, **56**, 472–477. (b)

Sarason, I. G. Relationships of measures of anxiety and experimental instructions to word association test performance. *Journal of Abnormal and Social Psychology*, 1959, **59**, 37–42.

Sarason, I. G. Empirical findings and theoretical problems in the use of anxiety scales. *Psychological Bulletin*, 1960, **57**, 403–415.

Sarason, I. G. The effects of anxiety and threat on solution of a difficult task. *Journal of Abnormal and Social Psychology*, 1961, **62**, 165–168.

Sarason, I. G. Verbal learning, modeling, and juvenile delinquency. *American Psychologist*, 1968, **4**, 254–266.

Sarason, I. G., & Ganzer, V. J. Anxiety, reinforcement, and experimental instructions in a free verbalization situation. *Journal of Abnormal and Social Psychology*, 1962, **65**, 300–307.

Sarason, I. G., & Ganzer, V. J. Effects of test anxiety and reinforcement history on verbal behavior. *Journal of Abnormal and Social Psychology*, 1963, **67**, 513–519.

Sarason, I. G., & Harmatz, M. G. Sex differences and experimental conditions in serial learning. *Journal of Personality and Social Psychology*, 1965, **1**, 521–524.

Sarason, I. G., & Koenig, K. P. The relationship of test anxiety and hostility to description of self and parents. *Journal of Personality and Social Psychology*, 1965, **2**, 617–621.

Sarason, I. G., & Minard, J. Test anxiety, experimental instructions and the Wechsler Adult Intelligence Scale. *Journal of Educational Psychology*, 1962, **53**, 299–302.

Sarason, I. G., & Palola, E. G. The relationship of test and general anxiety, difficulty of task, and experimental instructions to performance. *Journal of Experimental Psychology*, 1960, **59**, 185–191.

Sarason, I. G., Pederson, A. M., & Nyman, B. Test anxiety and the observation of models. *Journal of Personality*, 1968, **36**, 493–511.

Sarason, S. B., Davidson, K. S., Lighthall, F. F., Waite, R. R., & Ruebush, B. K. *Anxiety in elementary school children*. New York: Wiley, 1960.

Sarason, S. B., Mandler, G., & Craighill, P. G. The effect of differential instructions on anxiety and learning. *Journal of Abnormal and Social Psychology*, 1952, **47**, 561–565.

Sieber, J. E. A paradigm for experimental modification of the effects of test anxiety on cognitive processes. *American Educational Research Journal*, 1969, **6**, 46–61.

Silverman, R. E. Anxiety and the mode of response. *Journal of Abnormal and Social Psychology*, 1954, **49**, 538–542.

SILVERMAN, R. E., & BLITZ, B. Learning and two kinds of anxiety. *Journal of Abnormal and Social Psychology,* 1956, **52**, 301–303.

SNIDER, J. G., & OETTING, E. R. Autogenic training and the treatment of examination anxiety in students. *Journal of Clinical Psychology,* 1966, **22**, 111–114.

SPIEGLER, M. D., MORRIS, L. W., & LIEBERT, R. M. Cognitive and emotional components of test anxiety: Temporal factors. *Psychological Reports,* 1968, **22**, 451–456.

STANFORD, D., DEMBER, W., & STANFORD, F. A children's form of the Alpert-Haber achievement anxiety scale. *Child Development,* 1963, **34**, 1027–1032.

SUINN, R. M. The desensitization of test-anxiety by group and individual treatment. *Behaviour Research and Therapy,* 1968, **6**, 385–387.

TAYLOR, J. A. A personality scale of manifest anxiety. *Journal of Abnormal and Social Psychology,* 1953, **48**, 285–290.

TECCE, J. J., & HAPP, S. Effects of shock-arousal on a card-sorting test of color-word interference. *Perceptual and Motor Skills,* 1964, **19**, 905–906.

TRAPP, E. P., & KAUSLER, P. H. Test anxiety level and goal-setting behavior. *Journal of Consulting Psychology,* 1958, **22**, 31–34.

TREISMAN, A. M. Selective attention in man. *British Medical Bulletin,* 1964, **20**, 12–16.

WACHTEL, P. L. Anxiety, cue utilization, and patterns of ego defense. *Dissertation Abstracts,* 1966, **27**, 2149.

WACHTEL, P. L. Anxiety, cue utilization, and patterns of ego defense. (Doctoral dissertation, Yale University) Ann Arbor, Mich.: University Microfilms, 1966. No. 66-5420.

WACHTEL, P. L. Anxiety, attention, and coping with threat. *Journal of Abnormal Psychology,* 1968, **73**, 137–143.

WEST, C. K., LEE, J. F., & ANDERSON, T. H. The influence of test anxiety on the selection of relevant from irrelevant information. *The Journal of Educational Research,* 1969, **63**, 51–52.

WINE, J. Investigations of an attentional interpretation of test anxiety. Unpublished doctoral dissertation, University of Waterloo, 1970.

WONDERLIC, E. F. *Wonderlic personnel tests.* Northfield, Illinois: E. F. Wonderlic and Associates, 1959.

ZAFFY, D., & BRUNING, J. L. Drive and range of cue utilization. *Journal of Experimental Psychology,* 1966, **71**, 382–384.

(Received February 10, 1970)

Psychological Bulletin
1971, Vol. 76, No. 2, 105–110

BELIEF IN THE LAW OF SMALL NUMBERS

AMOS TVERSKY AND DANIEL KAHNEMAN[1]

Hebrew University of Jerusalem

People have erroneous intuitions about the laws of chance. In particular, they regard a sample randomly drawn from a population as highly representative, that is, similar to the population in all essential characteristics. The prevalence of the belief and its unfortunate consequences for psychological research are illustrated by the responses of professional psychologists to a questionnaire concerning research decisions.

"Suppose you have run an experiment on 20 subjects, and have obtained a significant result which confirms your theory ($z = 2.23$, $p < .05$, two-tailed). You now have cause to run an additional group of 10 subjects. What do you think the probability is that the results will be significant, by a one-tailed test, separately for this group?"

If you feel that the probability is somewhere around .85, you may be pleased to know that you belong to a majority group. Indeed, that was the median answer of two small groups who were kind enough to respond to a questionnaire distributed at meetings of the Mathematical Psychology Group and of the American Psychological Association.

On the other hand, if you feel that the probability is around .48, you belong to a minority. Only 9 of our 84 respondents gave answers between .40 and .60. However, .48 happens to be a much more reasonable estimate than .85.[2]

Apparently, most psychologists have an exaggerated belief in the likelihood of successfully replicating an obtained finding. The sources of such beliefs, and their consequences for the conduct of scientific inquiry, are what this paper is about. Our thesis is that people have strong intuitions about random sampling; that these intuitions are wrong in fundamental respects; that these intuitions are shared by naive subjects and by trained scientists; and that they are applied with unfortunate consequences in the course of scientific inquiry.

We submit that people view a sample randomly drawn from a population as highly representative, that is, similar to the population in all essential characteristics. Consequently, they expect any two samples drawn from a particular population to be more similar to one another and to the population than sampling theory predicts, at least for small samples.

The tendency to regard a sample as a representation is manifest in a wide variety of situations. When subjects are instructed to generate a random sequence of hypothetical tosses of a fair coin, for example, they produce sequences where the proportion of

---

[1] The ordering of authors is random. We wish to thank the many friends and colleagues who commented on an earlier version, and in particular we are indebted to Maya Bar-Hillel, Jack Block, Jacob Cohen, Louis L. Guttman, John W. Tukey, Ester Samuel, and Gideon Shwarz.

Requests for reprints should be sent to Amos Tversky, Center for Advanced Study in the Behavioral Sciences, 202 Junipero Serra Boulevard, Stanford, California 94305.

[2] The required estimate can be interpreted in several ways. One possible approach is to follow common research practice, where a value obtained in one study is taken to define a plausible alternative to the null hypothesis. The probability requested in the question can then be interpreted as the power of the second test (i.e., the probability of obtaining a significant result in the second sample) against the alternative hypothesis defined by the result of the first sample. In the special case of a test of a mean with known variance, one would compute the power of the test against the hypothesis that the population mean equals the mean of the first sample. Since the size of the second sample is half that of the first, the computed probability of obtaining $z \geq 1.645$ is only .473. A theoretically more justifiable approach is to interpret the requested probability within a Bayesian framework and compute it relative to some appropriately selected prior distribution. Assuming a uniform prior, the desired posterior probability is .478. Clearly, if the prior distribution favors the null hypothesis, as is often the case, the posterior probability will be even smaller.

heads in any short segment stays far closer to .50 than the laws of chance would predict (Tune, 1964). Thus, each segment of the response sequence is highly representative of the "fairness" of the coin. Similar effects are observed when subjects successively predict events in a randomly generated series, as in probability learning experiments (Estes, 1964) or in other sequential games of chance. Subjects act as if *every* segment of the random sequence must reflect the true proportion: if the sequence has strayed from the population proportion, a corrective bias in the other direction is expected. This has been called the gambler's fallacy.

The heart of the gambler's fallacy is a misconception of the fairness of the laws of chance. The gambler feels that the fairness of the coin entitles him to expect that any deviation in one direction will soon be cancelled by a corresponding deviation in the other. Even the fairest of coins, however, given the limitations of its memory and moral sense, cannot be as fair as the gambler expects it to be. This fallacy is not unique to gamblers. Consider the following example:

> The mean IQ of the population of eighth graders in a city is *known* to be 100. You have selected a random sample of 50 children for a study of educational achievements. The first child tested has an IQ of 150. What do you expect the mean IQ to be for the whole sample?

The correct answer is 101. A surprisingly large number of people believe that the expected IQ for the sample is still 100. This expectation can be justified only by the belief that a random process is self-correcting. Idioms such as "errors cancel each other out" reflect the image of an active self-correcting process. Some familiar processes in nature obey such laws: a deviation from a stable equilibrium produces a force that restores the equilibrium. The laws of chance, in contrast, do not work that way: deviations are not canceled as sampling proceeds, they are merely diluted.

Thus far, we have attempted to describe two related intuitions about chance. We proposed a representation hypothesis according to which people believe samples to be very similar to one another and to the population from which they are drawn. We also suggested that people believe sampling to be a self-correcting process. The two beliefs lead to the same consequences. Both generate expectations about characteristics of samples, and the variability of these expectations is less than the true variability, at least for small samples.

The law of large numbers guarantees that very large samples will indeed be highly representative of the population from which they are drawn. If, in addition, a self-corrective tendency is at work, then small samples should also be highly representative and similar to one another. People's intuitions about random sampling appear to satisfy the law of small numbers, which asserts that the law of large numbers applies to small numbers as well.

Consider a hypothetical scientist who lives by the law of small numbers. How would his belief affect his scientific work? Assume our scientist studies phenomena whose magnitude is small relative to uncontrolled variability; that is, the signal-to-noise ratio in the messages he receives from nature is low. Our scientist could be a meteorologist, a pharmacologist, or perhaps a psychologist.

If he believes in the law of small numbers, the scientist will have exaggerated confidence in the validity of conclusions based on small samples. To illustrate, suppose he is engaged in studying which of two toys infants will prefer to play with. Of the first five infants studied, four have shown a preference for the same toy. Many a psychologist will feel some confidence at this point, that the null hypothesis of no preference is false. Fortunately, such a conviction is not a sufficient condition for journal publication, although it may do for a book. By a quick computation, our psychologist will discover that the probability of a result as extreme as the one obtained is as high as 3/8 under the null hypothesis.

To be sure, the application of statistical hypothesis testing to scientific inference is beset with serious difficulties. Nevertheless, the computation of significance levels (or likelihood ratios, as a Bayesian might prefer) forces the scientist to evaluate the obtained effect in terms of a *valid* estimate of sampling variance rather than in terms of his subjective biased estimate. Statistical tests, therefore, protect the scientific community against overly hasty rejections of the null hypothesis (i.e., Type I error) by policing its many members who would rather live by the law of small numbers. On the other hand, there are no comparable safeguards against the risk of failing to confirm a valid research hypothesis (i.e., Type II error).

Imagine a psychologist who studies the correlation between need for Achievement and grades. When deciding on sample size, he may reason as follows: "What correlation do I expect? $r = .35$. What N do I need to make the result significant? (Looks at table.) $N = 33$. Fine, that's my sample." The only flaw in this reasoning is that our psychologist has forgotten about sampling variation, possibly because he believes that any sample must be highly representative of its population. However, if his guess about the correlation in the population is correct, the correlation in the sample is about as likely to lie below or above .35. Hence, the likelihood of obtaining a significant result (i.e., the power of the test) for $N = 33$ is about .50.

In a detailed investigation of statistical power, Cohen (1962, 1969) has provided plausible definitions of large, medium, and small effects and an extensive set of computational aids to the estimation of power for a variety of statistical tests. In the normal test for a difference between two means, for example, a difference of $.25\sigma$ is small, a difference of $.50\sigma$ is medium, and a difference of 1σ is large, according to the proposed definitions. The mean IQ difference between clerical and semiskilled workers is a medium effect. In an ingenious study of research practice, Cohen (1962) reviewed all the statistical analyses published in one volume of the *Journal of Abnormal and Social Psychology*, and computed the likelihood of detecting each of the three sizes of effect. The average power was .18 for the detection of small effects, .48 for medium effects, and .83 for large effects. If psychologists typically expect medium effects and select sample size as in the above example, the power of their studies should indeed be about .50.

Cohen's analysis shows that the statistical power of many psychological studies is ridiculously low. This is a self-defeating practice: it makes for frustrated scientists and inefficient research. The investigator who tests a valid hypothesis but fails to obtain significant results cannot help but regard nature as untrustworthy or even hostile. Furthermore, as Overall (1969) has shown, the prevalence of studies deficient in statistical power is not only wasteful but actually pernicious: it results in a large proportion of invalid rejections of the null hypothesis among published results.

Because considerations of statistical power are of particular importance in the design of replication studies, we probed attitudes concerning replication in our questionnaire.

> Suppose one of your doctoral students has completed a difficult and time-consuming experiment on 40 animals. He has scored and analyzed a large number of variables. His results are generally inconclusive, but one before-after comparison yields a highly significant $t = 2.70$, which is surprising and could be of major theoretical significance.
>
> Considering the importance of the result, its surprisal value, and the number of analyses that your student has performed—
>
> Would you recommend that he replicate the study before publishing? If you recommend replication, how many animals would you urge him to run?

Among the psychologists to whom we put these questions there was overwhelming sentiment favoring replication: it was recommended by 66 out of 75 respondents, probably because they suspected that the single significant result was due to chance. The median recommendation was for the doctoral student to run 20 subjects in a replication study. It is instructive to consider the likely consequences of this advice. If the mean and the variance in the second sample are actually identical to those in the first sample, then the resulting value of t will be 1.88. Following the reasoning of Footnote 2, the student's chance of obtaining a significant result in the replication is only slightly above one-half (for $p = .05$, one-tail test). Since we had anticipated that a replication sample of 20 would appear reasonable to our respondents, we added the following question:

> Assume that your unhappy student has in fact repeated the initial study with 20 additional animals, and has obtained an insignificant result in the same direction, $t = 1.24$. What would you recommend now? Check one: [the numbers in parentheses refer to the number of respondents who checked each answer]
>
> (*a*) He should pool the results and publish his conclusion as fact. (0)

> (*b*) He should report the results as a tentative finding. (26)
> (*c*) He should run another group of [median = 20] animals. (21)
> (*d*) He should try to find an explanation for the difference between the two groups. (30)

Note that regardless of one's confidence in the original finding, its credibility is surely enhanced by the replication. Not only is the experimental effect in the same direction in the two samples but the magnitude of the effect in the replication is fully two-thirds of that in the original study. In view of the sample size (20), which our respondents recommended, the replication was about as successful as one is entitled to expect. The distribution of responses, however, reflects continued skepticism concerning the student's finding following the recommended replication. This unhappy state of affairs is a typical consequence of insufficient statistical power.

In contrast to Responses *b* and *c*, which can be justified on some grounds, the most popular response, Response *d*, is indefensible. We doubt that the same answer would have been obtained if the respondents had realized that the difference between the two studies does not even approach significance. (If the variances of the two samples are equal, t for the difference is .53.) In the absence of a statistical test, our respondents followed the representation hypothesis: as the difference between the two samples was larger than they expected, they viewed it as worthy of explanation. However, the attempt to "find an explanation for the difference between the two groups" is in all probability an exercise in explaining noise.

Altogether our respondents evaluated the replication rather harshly. This follows from the representation hypothesis: if we expect all samples to be very similar to one another, then almost all replications of a valid hypothesis should be statistically significant. The harshness of the criterion for successful replication is manifest in the responses to the following question:

> An investigator has reported a result that you consider implausible. He ran 15 subjects, and reported a significant value, $t = 2.46$. Another investigator has attempted to duplicate his procedure, and he obtained a nonsignificant value of t with the same number of subjects. The direction was the same in both sets of data.
>
> You are reviewing the literature. What is the highest value of t in the second set of data that you would describe as a failure to replicate?

The majority of our respondents regarded $t = 1.70$ as a failure to replicate. If the data of two such studies ($t = 2.46$ and $t = 1.70$) are pooled, the value of t for the combined data is about 3.00 (assuming equal variances). Thus, we are faced with a paradoxical state of affairs, in which the same data that would increase our confidence in the finding when viewed as part of the original study, shake our confidence when viewed as an independent study. This double standard is particularly disturbing since, for many reasons, replications are usually considered as independent studies, and hypotheses are often evaluated by listing confirming and disconfirming reports.

Contrary to a widespread belief, a case can be made that a replication sample should often be larger than the original. The decision to replicate a once obtained finding often expresses a great fondness for that finding and a desire to see it accepted by a skeptical community. Since that community unreasonably demands that the replication be independently significant, or at least that it approach significance, one must run a large sample. To illustrate, if the unfortunate doctoral student whose thesis was discussed earlier assumes the validity of his initial result ($t = 2.70$, $N = 40$), and if he is willing to accept a risk of only .10 of obtaining a t lower than 1.70, he should run approximately 50 animals in his replication study. With a somewhat weaker initial result ($t = 2.20$, $N = 40$), the size of the replication sample required for the same power rises to about 75.

That the effects discussed thus far are not limited to hypotheses about means and variances is demonstrated by the responses to the following question:

> You have run a correlational study, scoring 20 variables on 100 subjects. Twenty-seven of the 190 correlation coefficients are significant at the .05 level; and 9 of these are significant beyond the .01 level. The mean absolute level of the significant correlations is .31, and the pattern of results is very reasonable on theoretical grounds. How many of the 27

significant correlations would you expect to be significant again, in an exact replication of the study, with $N = 40$?

With $N = 40$, a correlation of about .31 is required for significance at the .05 level. This is the mean of the significant correlations in the original study. Thus, only about half of the originally significant correlations (i.e., 13 or 14) would remain significant with $N = 40$. In addition, of course, the correlations in the replication are bound to differ from those in the original study. Hence, by regression effects, the initially significant coefficients are most likely to be reduced. Thus, 8 to 10 repeated significant correlations from the original 27 is probably a generous estimate of what one is entitled to expect. The median estimate of our respondents is 18. This is more than the number of repeated significant correlations that will be found if the correlations are recomputed for 40 subjects randomly selected from the original 100! Apparently, people expect more than a mere duplication of the original statistics in the replication sample; they expect a duplication of the significance of results, with little regard for sample size. This expectation requires a ludicrous extension of the representation hypothesis; even the law of small numbers is incapable of generating such a result.

The expectation that patterns of results are replicable almost in their entirety provides the rationale for a common, though much deplored practice. The investigator who computes all correlations between three indexes of anxiety and three indexes of dependency will often report and interpret with great confidence the single significant correlation obtained.[3] His confidence in the shaky finding stems from his belief that the obtained correlation matrix is highly representative and readily replicable.

In review, we have seen that the believer in the law of small numbers practices science as follows:

- He gambles his research hypotheses on small samples without realizing that the odds against him are unreasonably high. He overestimates power.
- He has undue confidence in early trends (e.g., the data of the first few subjects) and in the stability of observed patterns (e.g., the number and identity of significant results). He overestimates significance.
- In evaluating replications, his or others', he has unreasonably high expectations about the replicability of significant results. He underestimates the breadth of confidence intervals.
- He rarely attributes a deviation of results from expectations to sampling variability, because he finds a causal "explanation" for any discrepancy. Thus, he has little opportunity to recognize sampling variation in action. His belief in the law of small numbers, therefore, will forever remain intact.

Our questionnaire elicited considerable evidence for the prevalence of the belief in the law of small numbers.[4] Our typical respondent is a believer, regardless of the group to which he belongs. There were practically no differences between the median responses of audiences at a mathematical psychology meeting and at a general session of the American Psychological Association convention, although we make no claims for the representativeness of either sample. Apparently, acquaintance with formal logic and with probability theory does not extinguish erroneous intuitions. What, then, can be done? Can the belief in the law of small numbers be abolished or at least controlled?

Research experience is unlikely to help much, because sampling variation is all too easily "explained." Corrective experiences are those that provide neither motive nor opportunity for spurious explanation. Thus, a student in a statistics course may draw repeated samples of given size from a population, and learn the effect of sample size on sampling variability from personal observation. We are far from certain, however, that expectations can be corrected in this manner, since related biases, such as the gambler's fallacy, survive considerable contradictory evidence.

---

[3] Examples can be supplied on demand.

[4] Edwards (1968) has argued that people fail to extract sufficient information or certainty from probabilistic data; he called this failure conservatism. Our respondents can hardly be described as conservative. Rather, in accord with the representation hypothesis, they tend to extract more certainty from the data than the data, in fact, contain. The conditions under which people may appear conservative are discussed in Kahnemann, D., and Tversky, A. *Subjective probability: A judgment of representativeness.* (Tech. Rep.) Oregon Research Institute, 1971, 2 (2).

Even if the bias cannot be unlearned, students can learn to recognize its existence and take the necessary precautions. Since the teaching of statistics is not short on admonitions, a warning about biased statistical intuitions may not be out of place. The obvious precaution is computation. The believer in the law of small numbers has incorrect intuitions about significance level, power, and confidence intervals. Significance levels are usually computed and reported, but power and confidence limits are not. Perhaps they should be.

Explicit computation of power, relative to some reasonable hypothesis, for instance, Cohen's (1962, 1969) small, large, and medium effects, should surely be carried out before any study is done. Such computations will often lead to the realization that there is simply no point in running the study unless, for example, sample size is multiplied by four. We refuse to believe that a serious investigator will knowingly accept a .50 risk of failing to confirm a valid research hypothesis. In addition, computations of power are essential to the interpretation of negative results, that is, failures to reject the null hypothesis. Because readers' intuitive estimates of power are likely to be wrong, the publication of computed values does not appear to be a waste of either readers' time or journal space.

In the early psychological literature, the convention prevailed of reporting, for example, a sample mean as $\bar{X} \pm PE$, where PE is the probable error (i.e., the 50% confidence interval around the mean). This convention was later abandoned in favor of the hypothesis-testing formulation. A confidence interval, however, provides a useful index of sampling variability, and it is precisely this variability that we tend to underestimate. The emphasis on significance levels tends to obscure a fundamental distinction between the size of an effect and its statistical significance. Regardless of sample size, the size of an effect in one study is a reasonable estimate of the size of the effect in replication. In contrast, the estimated significance level in a replication depends critically on sample size. Unrealistic expectations concerning the replicability of significance levels may be corrected if the distinction between size and significance is clarified, and if the computed size of observed effects is routinely reported. From this point of view, at least, the acceptance of the hypothesis-testing model has not been an unmixed blessing for psychology.

The true believer in the law of small numbers commits his multitude of sins against the logic of statistical inference in good faith. The representation hypothesis describes a cognitive or perceptual bias, which operates regardless of motivational factors. Thus, while the hasty rejection of the null hypothesis is gratifying, the rejection of a cherished hypothesis is aggravating, yet the true believer is subject to both. His intuitive expectations are governed by a consistent misperception of the world rather than by opportunistic wishful thinking. Given some editorial prodding, he may be willing to regard his statistical intuitions with proper suspicion and replace impression formation by computation whenever possible.

REFERENCES

Cohen, J. The statistical power of abnormal-social psychological research. *Journal of Abnormal and Social Psychology*, 1962, **65**, 145–153.

Cohen, J. *Statistical power analysis in the behavioral sciences*. New York: Academic Press, 1969.

Edwards, W. Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment*. New York: Wiley, 1968.

Estes, W. K. Probability learning. In A. W. Melton (Ed.), *Categories of human learning*. New York: Academic Press, 1964.

Overall, J. E. Classical statistical hypothesis testing within the context of Bayesian theory. *Psychological Bulletin*, 1969, **71**, 285–292.

Tune, G. S. Response preferences: A review of some relevant literature. *Psychological Bulletin*, 1964, **61**, 286–302.

(Received February 26, 1970)

Psychological Bulletin
1971, Vol. 76, No. 2, 111–127

COLLEGE STUDENTS AS THERAPEUTIC AGENTS

GENE GARY GRUVER[1]

University of Arizona

The mutually beneficial effects of college students working as therapeutic agents with individuals in a variety of mental health settings are reviewed. Studies based primarily on observation and clinical impressionism suggest that college students may be useful as therapeutic agents, and, at the same time, students involved in a helping relationship exhibit personality changes not unlike those effected by more traditional psychotherapies. There is evidence that many unresolved issues will require further empirical investigation before the widespread use of nonprofessionals in psychotherapeutic settings can be sanctioned.

Albee's continuing investigation (1959, 1967, 1968) of professional manpower resources in the mental health fields has made apparent the severe current shortages and, further, has suggested the probability of even greater future shortages. It has been established that with present methods and models in the mental health field it is simply not possible to satisfy the ever increasing social demands for professional personnel.

A second aspect of the mental health manpower dilemma is that even if there were sufficient numbers of professional personnel, present mental health ideology would prevent many of those needing help from receiving it (Arnhoff, Rubinstein, & Speisman, 1969). This fact is most evident in the notorious lack of success mental health professionals have had working with the poor (Grosser, Henry, & Kelly, 1969; Pearl & Riessman, 1965). Other groups such as drug abusers, alcoholics, and juvenile delinquents have also been neglected by the mental health professionals primarily because professional contact with them has been, for the most part, fruitless. On the other hand, Blau (1969b) reported that nonprofessionals appear to have had a great deal of success in working with these groups. Particularly effective have been programs such as Alcoholics Anonymous and Synanon (Volkman & Cressey, 1964; Yablonsky & Dederich, 1963).

In addition to the success of nonprofessionals working directly with these individuals, another kind of role for the nonprofessional may be useful to mental health conceptualization. Reiff (1967) and Pearl and Riessman (1965) have suggested that one of the most important uses of nonprofessional personnel is a "bridging" function, that is, a link between professionals and their target populations. In order to be an effective "bridge," the nonprofessional must be indigenous of the target population and speak the vernacular of both the target population and the professional. His interclass communication and mediation skills can facilitate communication between the two groups, and, thus, he will be able to discuss the needs and goals of each group with the other.

Other groups go untreated, not because mental health professionals are less successful in treating them, but rather because professionals choose situations that are more challenging and may prefer to utilize their time working with individuals who manifest dramatic change more rapidly. Nonprofessionals may be able to fill this void by working effectively with some of these neglected groups, such as retarded children and geriatrics. In this way, nonprofessional mental health personnel may be the key to providing more complete coverage in terms of mental health services to the entire population.

More provocative is the evidence that some nonprofessionals are more effective than their professional counterparts in working with some populations which are presently receiving professional focus. Zunker and Brown (1966), for example, found that student counselors were more effective in counseling college students than were the professional counselors. Poser's (1966) college girls produced more positive changes in hospitalized chronic schizophrenic patients than did the professional staff. If indeed nonprofessionals are more effective than professionals in dealing with certain populations, then part of our professional mental health manpower resources is being wasted!

---

[1] An earlier draft of this paper was submitted to the University of Arizona in partial fulfillment of the preliminary requirements for the PhD degree. I wish to thank Peter Madison for his direction and suggestions. I also wish to thank R. Wrenn, R. A. Ruiz, C. Trafton, and R. Pool for their helpful comments.

Requests for reprints should be sent to Gene Gary Gruver, 731 North Ruston Avenue, Tucson, Arizona 85711.

It is necessary to take a new look at traditional conceptualizations in mental health, to examine the parameters of professional involvement in social problems, and to set goals for the future. All aspects of the manpower dilemma must be investigated to insure maximum utilization of present resources as well as seeking out new sources of mental workers.

Attempts at amelioration of the manpower dilemma in psychology have included subdoctoral education (Albee, 1968; Arnhoff & Jenkins, 1969; Hoch, Ross, & Winder, 1966; Knott, 1969), the examining of new models for psychology (Albee, 1968; Hoch et al., 1966), and the utilization of volunteers or nonprofessionals (Christmas, 1966; Holand & Voss, 1968; Reding & Goldsmith, 1967; Savino & Schlamp, 1967; Wahler, Winkel, Peterson, & Morrison, 1965).

Volunteers or nonprofessionals with varying levels of training, motivation, and socioeconomic status have been used in a variety of situations. Two investigations by Stover and Guerney (1967) and Patterson and Brodsky (1966) have utilized mothers as therapeutic agents for their own children while Rioch, Elkes, Flint, Usdansky, Newman, and Silber (1963), Nichtern, Donahue, O'Shea, Marans, Curtis, and Brody (1964), and Donahue and Nichtern (1965) have trained mature women to act as therapists for the children of others. Hirsch (1968) and Zimmerman and Zimmerman (1962) have shown that teachers can be effective as therapeutic agents. Even grandparents and retired individuals have been put to work as psychotherapeutic agents (Cowen, Leibowitz, & Leibowitz, 1968; Johnston, 1967). Hartog (1967) has described the work of volunteer enlisted men consultants in a United States Army hospital, while Fellows and Wolpin (1969) and Perlmutter and Durham (1965) have described the use of high school teenagers in mental health programs. Grosser et al. (1969) have reported on the Conference on Nonprofessionals, which dealt primarily with the use of indigenous personnel in health and welfare fields.

The present review centers around college students who act as volunteers or nonprofessionals in various mental health facilities. The focus upon college students is relevant not only because they appear to have qualities and characteristics which may suit them to work with troubled individuals but also because such work appears to have a positive developmental influence upon their own personalities. Because of gross methodological deficiencies in most of the studies involving college students as therapeutic agents, these impressions and this review are based primarily upon conceptual analysis of present day mental health issues, its logic, and clinical observation. The justification of nonprofessionals and the issue of college student development need more study before widespread use of college students in mental health settings can be sanctioned. The function of this review, then, is to identify and structure important areas which require further empirical investigation.

Observers have postulated many unique characteristics of college students which make them especially desirable as mental health workers. Greenblatt and Kantor (1962a) suggested that college students are more successful than "volunteers of a more senior station in life," that is, ladies auxiliary and the traditional friends of mental hospitals, because they manifest less resistance to and more motivation for face-to-face contact with patients. Whereas more mature adults tend to hesitate at the door of the ward and restrict their service by selecting duties that keep them some distance from the patients, the college students courageously plunge into the ward and usually are able to make direct long-term contact with even the very severely regressed patients. Also, student volunteers appear to have a sense of personal conviction to their work that the staff of other volunteer workers cannot duplicate. Umbarger, Dalsimer, Morrison, and Breggin (1962) suggested that the reason for student success with patients results primarily from three factors. First of all, they are crusaders who feel the exhilaration of finding a worthwhile cause. Further, they are revolutionists engaged in a struggle against mental illness, the toll of which can be seen in mental hospitals. Finally, an element of altruistic novelty characterized the students observed by Umbarger et al. (1962).

Mitchell (1966) posited that in working with children "college students seem to have a particular talent for finding the child in his own world [p. 311]." He further suggested that the casual dress of the college students, their idealism, and, in particular, their spontaneity and enthusiasm are all disarming features which enhance the effectiveness of their role. In his argument against equating behavior of college girls in mental hospitals with psychotherapy, Rosenbaum (1966) reasoned that college students, particularly young girls, are successful because mental hospital populations who have been rejected by our culture "will respond to young, vibrant people who are humane and extend warmth [p. 294]."

Similarities such as residing in a developmental institution and seeking life changes may create a bond between college students and inmates of mental hospitals and constitute a further basis for the reported success of college students working with troubled individuals. Keniston (1967) pointed out that both college students and psychotic patients reside in "developmental institutions," the common characteristics of which include stimulating, supporting, and confirming the development of each resident (college students would certainly see even more similarities between the two institutions). Sanford (1962) also compared college students to mental hospital patients in that they are both seeking personality change.

The fact that there is a reduction of social distance between college students and those seeking help may facilitate the establishment of a working relationship between the therapeutic agent and patient. Certainly, the distressed patient might expect more empathy from a college student who, like himself, is struggling with his identity, competing for financial and employment security, and who also sees the locus of control outside himself. The professional therapist, on the other hand, may appear as though he long since attained identity, security, and control of self and thus cannot recall the intensity of his own struggle. If then, as some authors suggest (Rogers, 1957; Truax & Wargo, 1966), accurate empathy is the single most important therapist characteristic in determining outcome of therapy, then the college student instantly and automatically has an advantage over his professional counterparts via his social status. Furthermore, if feeling closer to the therapist in terms of social status does facilitate "instant felt empathy," then college students should be even more effective working with other college students, adolescents, and children.

Because they cannot rely on professional training or the professional facade, college students are forced to use a naive, commonsense approach to their encounter. Perhaps the fresh approach of these young people has an effect on depressed or disturbed individuals that the trained professionals have difficulty duplicating. Rogers (1957) has postulated that theoretical wisdom is not a necessary ingredient for a helping relationship, and, to expand Rogers' contention, it may be that theoretical wisdom is actually detrimental to some therapeutic encounters because it restrains and constricts the variety of approaches available to the therapist. College students, on the other hand, are free of this theoretical constraint and are less inhibited in trying new approaches. As a result of their ignorance they may uncover effective new approaches considered inappropriate or too illogical by professionals. The present author is not advocating disregarding all theory but simply suggesting that we may be able to learn from the fresh approach of untrained college students.

Reiff and Riessman (1965) have postulated that greater flexibility in terms of appropriate and accepted behaviors on the part of the nonprofessional may be a special asset. Whereas a college student may take his charge to the zoo or be invited to a party by the patient, the professional, by virtue of his role prescription, would not engage in these activities. Poser's (1966) college girls could play "Drop-the-

Hanky" with their patients, but the professionals in the study would not have dared such activities.

A final advantage of using college students as therapeutic agents and a further reason for their apparent success particularly in settings outside the hospital is that there may be less stigma involved for the client. Whereas a parent may be concerned about the stigma attached to sending his child to a psychologist, he may be less inhibited about his child seeing a college student given, of course, that therapeutic effect is roughly similar. To sum, it appears that college students may have an advantage in working with some populations by virtue of their *not* having prior training or professional status.

Personality theorists who are particularly interested in college student development (Madison, 1969; Sanford, 1962) suggest that college students have a significant potential for change, and there is a continuing search for ways in which the personality development can be facilitated. Evidence suggests that working part time in a community mental health facility may serve as an instrument of personality change. College students who work in mental hospitals, psychological clinics, or other mental health settings manifest significantly more positive changes in self-acceptance and moral judgments in sexual and aggressive acts than do control groups (Holzberg, Gewirtz, & Ebner, 1964) and also greater self-understanding (Reinherz, 1962; Stollak, 1969; Umbarger et al., 1962). Increased self-confidence and enhanced identity formation are further personality changes effected by working in a mental health installation (Scheibe, 1965; Umbarger et al., 1962).

Even if college students possessed none of the above characteristics which make them particularly desirable as therapeutic agents, the discovery of a large untapped pool of volunteer workers in mental health would be welcome news. The suggestion that they do have characteristics and qualities which make them particularly suited to working in a mental health setting, as well as evidence that working in mental health may foster desirable personality changes, makes the topic not only more interesting but also worthy of theoretical attention and of a review.

COMPANION PROGRAMS

Companion Program is a term used to describe situations in which college students spend a certain amount of time each week as "companions" to patients in mental hospitals. Companion Programs may be structured or unstructured, provide training or no training, provide monetary remuneration or no monetary remuneration, but they share the common feature that individuals from the community are brought into regular face-to-face contact with individuals evidencing behavior problems.

The first Companion Program originated in 1954 at Harvard University and provided service to the Metropolitan State Hospital (Umbarger et al., 1962). By 1962, the idea had spread to such an extent that a book describing one program was published (Umbarger et al., 1962), and a conference was held to discuss the impact of the programs (Kantor, 1962). The Companion Program idea is continuing to expand and is usually met with enthusiastic response from students, staff, and patients alike (Spoerl, 1968).

The program at Metropolitan State Hospital has been the model for subsequent Companion Programs, although not all new programs have been so ambitious as the original in which 2,000 students had been in contact with mental patients during the first 7 years of operation. At Metropolitan State Hospital there are four separate areas in which the students might choose to work. Each situation is formulated to bring students and patients together under fundamentally different conditions. First, there is the ward improvement project in which a group of students go onto a ward with severely regressed patients to paint the ward, hang paintings, hold parties, but most importantly, interact with patients. A second area of work is the children's unit, which includes both group and individual activities. Organization of a basketball team, academic tutoring or just going for walks are among the services performed by volunteers on the children's unit. The third area is the case-aide program in which students work in a one-to-one relationship with individual patients under professional supervision and control. Six years after the original program had begun, the fourth work area was initiated. Wellmet, Inc., is a halfway house for patients in transition between the hospital and the outside community. All of these four programs at Metropolitan State Hospital are loosely structured, provide training in the form of group meetings held once a week during the time students are entering services, and provide no monetary remuneration.

Using whatever measure one chooses, the program at Metropolitan State Hospital has been an overwhelming success. The effects on patients and student volunteers alike have been, for the most part, positive. Measures of success on the ward improvement projects were subjective and consisted of the students' accounts, which were written in the *Adult Unit Diary* and also included the recording of "landmarks," which were considered important advances in patient group behavior. One of these landmarks, for example, was the opening of doors between the women's "violent" ward and its male counterpart.

The measure of success in the case-aide program is not only more objective but more startling. By the end of the first year, 11 of the 14 patients who had been visited by case-aide volunteers had been released from the hospital with a readmission rate of less than 28%. Improvements in the other three patients could be measured by their transfers to less secure wards and by other idiosyncratic achievements.

In a follow-up study of the chronic psychotic patients seen by college case-aide volunteers, Beck, Kantor, and Gelineau (1963) found that of the 120 case-aide patients who had been seen from 1954 to 1961, 37 (or 31%) had left the hospital while working with students. Of these 37, 28 were still out of the hospital at the time of the follow-up, an average of 3.4 years after they left the hospital. Seven more patients left the hospital a few months after their case-aide work and were all out at follow-up, an average of 1.2 years each. The finding that 31% of a group of psychotics were able to leave a chronic service after "treatment" by college students appears to support the assertion that the program at Metropolitan State is successful, but a firm conclusion is not possible without controls.

As has been suggested, the measures of success of the program at Metropolitan State Hospital include not only the effects upon the patients but also the effects upon the student volunteers. All students in the program claimed "that they learned a great deal from the case-aide experience." Many felt that their relationship with a patient and the instruction of the group leader had taught them more about psychological theory and mental illness than had their courses at college. Some became interested in careers in mental health work. Moreover, "all claimed that they had gained insight into their own personalities and problems through their relationships with the patients and their own group [Umbarger et al., 1962, p. 54]."

Erikson (1959) suggested that the crystallization of professional goals is a major phase of the process identity formation. If this is true, then work in the case-aide section of the program can be said to have facilitated identity formation since Kantor (1959) and Greenblatt and Kantor (1962b) have shown that more than 70% of the students who were indefinite or undecided about career choices before participating in the case-aide program made concrete choices in the direction of mental health. In evaluating Kantor's findings, it should be noted that no control groups were used, and, moreover, students' career decisions during college years tend to be unstable. However, Kantor's essential conclusion that the project influenced the career choice of participants in the direction of mental health is probably valid.

As in most new areas of research, objective measures of college student success with patients are scarce; however, a large body of relatively objective research, particularly concerning the student development aspect, has come out of the Connecticut Valley Companion Program, which was modeled after the one at Metropolitan State. In one of the studies (Holzberg & Gewirtz, 1963), a group of students who volunteered for the companionship program were compared to a control group of students who volunteered for other social service activities such as YMCA or the Big Brothers. On a questionnaire that was administered to both groups at the beginning of the academic year and again at the conclusion of that year, the subjects in the Companion Program shifted significantly in a positive direction in terms of their attitudes toward and knowledge of mental illness.

In another study at Connecticut Valley Hospital, a questionnaire measure yielded data suggesting positive effects on both patients and students (Holzberg, Whiting, & Lowy, 1964). Eighty-four percent of the patients said they enjoyed the relationship with the students, while the students reported that 71% of the patients showed improvement over the year. Ninety-one percent of the students themselves reported they became less anxious about working in a mental hospital, 90% reported a greater understanding of mental illness, 84% suggested feelings about mental hospital personnel had changed, and 97% of the students considered that their experiences had contributed to their personal growth. In another study (Holzberg et al., 1964), the effects of association with hospitalized mental patients on the personalities of 32 male college students were compared to a control group of 24 students who had not been involved with mental patients. Students in the Companion Program demonstrated significantly positive change in self-acceptance and in moral judgments concerning sexual and aggressive behaviors. Holzberg and Knapp (1965) have presented further evidence of positive effects upon Companions in their findings that after serving as Companions they are less frequently on academic probation and that they increase their introspective behavior.

More recently Holzberg, Knapp, and Turner (1967) have collected psychological test data so as to compare patients in the Companion Program with a group of control patients who were not. A relatively small sample of 13 Companion patients and 30 controls completed the Minnesota Multiphasic Personality Inventory (MMPI), Bender Gestaldt Designs, and Draw-a-Person before and after the program. Companion patients changed significantly on the Depression scale of the MMPI from pretest to posttest, and a similar difference on the Paranoid scale barely missed significance. The investigators are still studying data on the Bender Gestaldt Designs and Draw-a-Person to see if significant changes occurred in Companion patients on those instruments.

One may argue that college students volunteering for service as a Companion appear to change to a greater degree because they are more pathological at the time they volunteer than the average college student. Knapp and Holzberg (1964) have shown this not to be the case. A group of 85 male college students volunteering for service as Companions to chronically ill mental patients were compared to a group of 85 control students on a number of psychological tests administered during the students' freshman year. Tests included MMPI, Edwards Personal Preference Schedule, Allport-Vernon-Lindzey Scale of Values, the Scholastic Aptitude Test, and the Terman Concept Mastery Test. Students taking part in the Companion Program were shown not to be different from their control counterparts in any significant clinical respect, but there is evidence that the Companions are (*a*) slightly more religiously oriented, (*b*) more morally concerned, (*c*) more compassionate, and (*d*) more introverted than the control students. In a later study, volunteers were also differentiated from controls on academic variables such as their major area of study, frequency of disciplinary action, and fraternity affiliations (Holzberg et al., 1967).

Scheibe (1965) described a program which is similar to the Companion Program model except that students were assigned to work for a continuous 8-week period rather than once-a-week appearances over a year's time as required in the Companion Program. Students in the Service Corps Program of the State of Connecticut lived at the hospital and spent a normal working week with chronic patients for which they received $200 salary for the 2-month period. Students were not assigned to a specific patient but worked with all the patients on the ward in unstructured activities. Positive changes in the students' description of the typical mental patient were noted on Gough's Adjective Check List given at the beginning and at the end of the work period. In describing themselves on the Gough Adjective Check List, students exhibited significant gains in Achievement, Dominance, Self-Confidence, and Nurturance. There were no reported adverse effects on the college students as a result of working with the mental patients. Further, Greenblatt and Kantor's (1962b) findings were substantiated in that a crystallization of vocational goals appeared in a direction favorable to mental health.

Hersch, Kulik, and Scheibe (1969) subsequently published a more detailed study of personal characteristics of college volunteers in the Service Corps Program. One hundred fifty-one students serving in the Connecticut Service Corps and 142 controls enrolled in summer school at four Connecticut colleges were given a battery of tests and questionnaires including California Psychological Inventory, Gough Adjective Check List, Rotter Internal–External Small Scale, Marlowe-Crowne Social Desirability Scale, the Strong Vocational Interest Blank, and a biographical questionnaire. The striking personal characteristics of the college student volunteers were maturity and control, drive for independent achievement, and sensitivity to distressed individuals. On the Strong Vocational Interest Blank, their interests were similar to those in professions emphasizing social service. Autobiographical data further indicated that the college student volunteers were more service oriented and more dedicated to mental health service. The authors concluded that "data reported here suggest that participation in volunteer work is not motivated by overconcern with personal problems but rather is partly attributable to a controlled drive for independent achievement and sensitivity to human problems [p. 34]."

Lawton and Lipton (1963) reported on a project at Morristown State Hospital wherein six college students were employed full time. No training was given, but the students were instructed to devote all their efforts to creating highly personal relationships with individual patients. Based upon subjective reports of the students, positive changes were noted in individual patients.

Levine (1966) reported an investigation of the changes in attitude and behavior produced in students by a nonacademic, off-campus program which he suggested appeals to and puts to work the unenacted idealism of today's college youth. Recreational and social activities with the mental patients fostered more positive attitudes toward and increased interest in social action.

Poser's (1966) now classic study was designed to compare the effect of professional and untrained therapists. The untrained therapists were 11 college girls as well as two inpatients—none of whom had training in psychology. The professional therapists included psychiatrists, psychiatric social workers, and occupational therapists. Poser divided 343 male chronic schizophrenic patients into groups of 10 with each group matched as closely as possible with every other unit in terms of patient age, severity of illness, and length of hospitalization. Each of the groups were then assigned to a therapist picked at random. Each therapist met with his or her group for 1 hour each day, 5 days a week, over a period of 5 months. Both the trained and untrained therapists were free to conduct their therapy sessions as they wished. Results were interpreted using the difference in pretherapy and posttherapy scores on six different psychological tests. It was found that the college students achieved slightly better results than the professional mental health workers doing group therapy with similar patients.

Walker, Wolpin, and Fellows (1967) described a program which was a joint venture between Westmont College, Santa Barbara, and Camarillo State Hospital, Camarillo, California. Students received college credit for research and service activities involving direct contact with patients. Using a modified sentence-completion test for the students and subjective reports of the patients, they concluded that "we may be able to foster better personal developments as well as enrich school and college curricula while developing potential interest and entrance into the mental health field [p. 188]."

Spoerl (1968) reported on students from the University of Washington who volunteered to work on the psychiatric service of the University Hospital. In order to capitalize on the student spontaneity and imagination, no instructions were given except that the volunteers were to begin their relationship with patients on a peer basis. Volunteers could work with either individuals or groups. No objective evaluation of the student volunteer project on the patients is available, but questionnaires were given the ward staff and patients, and written reports were solicited from the volunteers. Most of the comments on the questionnaires and written reports were favorable to such an extent that the program has been enlarged to involve more volunteers and more patients.

Cowen (1969a) discussed an innovation in which college students are enrolled in a year-long seminar practicum in mental health. The first 2 months are used in examining current and past community mental health programs, and then the students, working in groups of six to eight each with a graduate student leader, begin service in one of three community settings: (*a*) a state hospital in which they encourage verbal and nonverbal interaction with chronic schizophrenic patients, (*b*) a school setting working with young students experiencing school adjustment problems, and (*c*) a settlement house with very young children who are exhibiting very slow social verbal development.

COLLEGE STUDENTS WITH CHILDREN

Many of the Companion Programs have special Children's Units, but, since children appear to be more malleable and become much more intensely involved with the student companion, volunteers for the Children's Units are usually subjected to a far greater amount of control than those working with adults. College students often prefer to work with children for a number of reasons. First, improvement in the younger patients is more easily observed even by naive volunteers. Also, students discover that in just a short time the children begin to respond positively to college students. Umbarger et al. (1962) reported that students working with the children felt less anxious about their own identity and more successful in their work than they did with the older patients. Students were apparently much more effective because they could act in a more relaxed and normal manner. Further, socially validated roles of big brother and big sister worked extremely well with the children while no such role was readily available with the adult patients.

Kreitzer (1969) used psychology majors who received course credit for their work with hospitalized emotionally disturbed children. Training consisted of completion of course work in psychopathology as well as a weekly 2-hour group supervision session. Measures included staff member ratings and teacher-rated changes in, or elimination of, inappropriate behavior. Many of the target behaviors were reduced or eliminated, and some of the student therapists involved called this the "highlight of their college experience."

Reinherz (1964) reported a project in which students from Radcliffe and Harvard volunteered for work at Massachusetts State Hospital working with children who were inpatients and who did not have severe behavioral problems. Volunteer college students spent one afternoon a week with the children after having met with a social work supervisor for 15 minutes ahead of time to receive a progress report of the child. During the first year of the program, ward psychiatrists reported improved functioning in three out of the four children in the program. In the second year, physicians reported change and progress in all seven patients, and, in several cases, psychological tests confirmed positive growth. At the end of the second year, two of the seven patients were ready for discharge, and a third had gone home on extended leave. Earlier, Reinherz (1962) had observed that some of the successes that college students have in working with emotionally disturbed children may be due to their having recently solved or left unsolved basic issues of maturation in their own lives. She noted that in late adolescence identity problems such as sex role and career choice are important developmental issues, and their successful resolution makes the difference between a productive and nonproductive adult role. Often it was observed that as the student aided the child in working out the problems of self-maturity, the student, too, appeared to be gaining a definitive solution for himself.

Fellows and Wolpin (1969) reported a project using teenage psychology trainees in a mental hospital. Four academically superior boys spent 4 days per week working with adolescent and preadolescent boys in a mental hospital. They were paid $50 per month for 2 months. Following a week of orientation at the hospital, the trainees began individual counseling on the adolescent male unit engaging in pilot conditioning procedures and conducting group therapy. Students were on their own except for 1 hour per day of permissive discussion with their supervisors. Measurement in the study was primarily subjective, with the authors reporting a particularly notable effect upon the patients in that the boys idealized the trainees and sought to copy whatever they did. More objectively, it is reported that a number of boys from this particular dormitory "have been placed in foster or group homes where they seem to be adjusting well, having learned some aspects of effective behavior from the trainees [p. 276]."

Heretofore the emphasis has been placed upon students working with an inpatient population in mental hospitals, but many day programs or outpatient programs have been initiated, particularly in caring for children. Belz, Drehmel, and Silvertsen (1967) suggested that volunteers are an essential ingredient in the ongoing functioning of a facility dedicated to the psychiatric treatment of children.

Brennan (1967) has shown that college students can be used as helpers in providing expanded mental health services for children. He pointed out that the attrition rate of children in guidance clinics has been found to be around 60% (Tuckman & Lavell, 1959). Although companionship may not replace certain diagnostic and treatment services, it does represent a service that parents and children will use. He based his suggestion on the low attrition rate of 12% for this particular study in which he used college students as Big Brothers. He concluded that college students may facilitate and enhance treatment of children with psychiatric problems.

Goodman (1967) has experimented with companionship therapy between college students and troubled boys. Male college students were trained in a 2½-day experimental workshop and were paid $1.40 per hour. After counselors were selected, they were divided into a "quiet" group an "outgoing" group, with half the quiet counselors being paired with boys evidencing quiet problems, that is, social introversion. The other half of the quiet counselors were paired with boys having outgoing problems, and the same procedure was followed for the group of outgoing counselors. Although only tentative findings are available, results suggest that boys with social introversion problems gain most from participating in the program. As in the Companion Programs, Goodman noted that his students manifested personality changes not unlike those reported by Holzberg (1963). Goodman's counselors showed a dramatic increase of interest in the behavior of children and in working with troubled people, and they reported that improvement occurred in the way they interact with friends. Differences between counselors and matched controls who did not participate in the Companion Program are significant. Nonparticipating controls were used so as to determine the effects on the counselor in working with children.

Mitchell (1966) has coined the term "amicatherapy" which he described "as a form of therapeutic intervention whereby layman volunteers relate in sustained friendship roles to troubled and disturbed persons under the guidance and supervision of professionals [p. 307]." He suggested that there may be advantages to using college students in amicatherapy since college students seem to have a particular knack for meeting the child in his own world. He postulated that this may be the case since students are not yet irrevocably committed to the adult world. After a student is chosen to work with a particular child, he sees a supervisor for an orientation interview and is instructed simply to make friends with the child. Clinical investigations of 74 children who participated in Mitchell's program have indicated "all of the children have benefited by their relationship with the student volunteers [p. 314]."

Stollak (1969) and Linden and Stollak (1969) have investigated the possibility of training college students as play therapists. In the former study, the students' role is modeled as closely as possible to that of a client-centered play therapist. The basic task is to be empathetic, understanding, nondirective, and to convey this understanding and acceptance to the child. Students were trained in 10 sessions during which they observed play therapy techniques and played with normal children. At the end of the tenth session, each student was assigned a child between the ages of 4 and 10 who were taken from the waiting list of the Lansing Child Guidance Clinic or the Psychology Clinic at Michigan State University. Stollak (1969) noted that undergraduates do significantly change their behavior during the sessions by increasing their reflection of content and clarification of feeling statements. Linden and Stollak (1969) concluded that communicated empathy is not an innate ability but must be taught. This has an important implication for the utilization of college students in mental health settings. If one adheres to the client-centered tenet that communication of accurate empathy is a necessary prerequisite for therapeutic movement, the turning loose of naive, untrained college students on a mental hospital is not as effective in producing therapeutic change as the same students might be if they were first taught to communicate the empathy they feel, that is, to make appropriate verbal statements.

Davison (1966) trained undergraduates as social reinforcers for autistic children at a private day care center; however, the results were less than significant when kind and degree of behavior change were used as the criteria for success. Less than 15 hours of the child's waking hours were spent in the program, and, since learning theory depends on environmental contingencies for factors that shape and maintain behavior, these results were not unexpected. One important finding which did emerge from this study, however, is that intelligent, motivated college students can be trained in a very short time to carry out a behavior control program that applies learning principles to the manipulation of psychotic behavior in children.

COLLEGE STUDENTS AS THERAPEUTIC AGENTS IN RELATED AREAS

Heretofore this study has focused upon college students as therapeutic agents in settings traditionally within the realm of mental health, that is, mental hospitals and psychological clinics. In recent years, however, mental health professionals have come to view other settings as legitimate surrounds for mental health workers. Cowen has headed a number of research efforts seeking prevention of emotional disorders in the school setting (Cowen, 1968; Cowen, 1969b; Cowen, Carlisle, & Kaufman, 1969; Cowen, Izzo, Miles, Telschow, Trost, & Zax, 1963; Cowen, Zax, Izzo, & Trost, 1966; Cowen, Zax, & Laird, 1966; Zax & Cowen, 1967). Cowen, Zax, and Laird (1966) selected 17 undergraduate volunteers to provide emotionally disturbed children with a meaningful relationship by pairing them with active, enthusiastic college students. Student volunteers had no training and were encouraged to foster a spontaneous, warm friendship with the child. For a variety of assumed reasons, none of the outcome measures discriminated between the experimental group of emotionally disturbed children. Certainly the short time the program was in effect (2 months) limited the number of contacts. Another reason there may have been no significant difference between the two groups is that the control group was simultaneously engaged in another program aimed at preventing emotional disturbance. There were significant changes among the volunteers, however, in that institutional concepts were no longer rated in a stereotypically positive way. Further, volunteers rated youngsters experiencing emotional difficulties in a more positive and accepting manner on a semantic differential. It will be recalled that similar changes toward more positive attitudes toward mental illness have been found to result from participation in Companion Programs (Holzberg & Gewirtz, 1963).

In another study, Cowen (1968) compared the effectiveness of two interventive programs —one using housewives, the other using college students. Both groups of volunteers worked with children experiencing emotional difficulties in a school setting. Two independent rating-scale evaluations were used to measure changes in behavior of the youngsters. Mean improvement scores of the two experimental groups combined were significantly greater than those of noncounseled control groups; however, only the group seen by the housewives was rated as significantly improved. Children seen by the college students were directionally better than controls, but the difference was not significant. Cowen postulated several reasons for this difference. First of all, the housewives were carefully screened, with the experimenter selecting only those applicants with attributes facilitating interpersonal interactions with young children. The college student selection was based only on gross negative screening which eliminated the most obviously unsuited. Also, the housewives had 2 years experience in the schools, whereas the college students had no previous experience. Further, the housewives knew school personnel better and were generally better integrated in the school setting. Also, the schedules of housewives were more flexible, and they were available at a greater number of times than were the college students. The subjective impressions of school personnel were that these facts stacked the cards in favor of the housewives.

J. McV. Hunt (1969) discussed a model for psychology he called the Hall-Nebraska "Model" where students are involved in a "counselor-counselee" relationship with various kinds of people who exhibit a variety of problems of living in the community. Undergraduate college students became college pals to deprived children, teenagers, families, children in orthopedic hospitals, children in institutions for emotional disorders, high school dropouts, and juvenile delinquents. Undergraduate college pals established an ongoing relationship with an individual in one of the aforementioned categories and continued his contact throughout the school year. When the counselor left school, he introduced to his counselee a new counselor and observed and encouraged the new relationship. The college pal association has been particularly effective in the family project. This program deals with 21 children—7 in each of three families, each child having a "college pal." There is no control group other than the children of other families in the neighborhood. Children from the neighborhoods of these three families seldom complete high school, and one criterion of the success of the college pal project was the number of the children in the project who have completed high school. All students in this project who are old enough to have completed high school have done so. Further, all have had at least a try at college. As with other projects of this kind, a second measure of success is the effect upon the college student themselves. Hunt reported that not only does this type of project keep counselors from dropping out of college, but also the students are learning about human relationships by dealing directly with people who are having problems in the community. Moreover, they appear to be learning an amazing amount about research methodology, including the vernacular.

The Department of Health, Education, and Welfare (1966) has noted that students represent a large and relatively untapped source of volunteer manpower for institutions housing juvenile delinquents. It is suggested that college students may more easily break the psychological barriers that often separate juvenile delinquents from adult workers since alienated young people usually trust another young person more than they trust adults. In the probation program at Boulder, Colorado, college students are presently serving as tutors, interviewers, and discussion leaders for the delinquents.

Gorlich (1967) further postulated that the functions of student volunteers in institutions for delinquents are threefold. First, they prove to the delinquent that someone on the outside really cares about them. Also, college students provide the young person with a role model. Finally, the students can later help spread the word about institutional needs.

Zunker and Brown (1966) supported the idea that youth-to-youth counseling may be more effective than adult-to-youth counseling. A sample of 106 college freshman, half males and half females, received $6\frac{1}{2}$ hours of academic adjustment guidance from same-sexed professional counselors. Trained under identical conditions, upperclassman student counselors gave equivalent guidance to all other beginning freshmen at Southwestern State Texas College. Student counselors also used identical guidance materials, followed identical counseling activity sequences, and were provided facilities equivalent to those used by the professional counselors. A matching sample of 80 men with 80 women was subsequently drawn from the freshmen receiving student-to-student counseling. Student counselors were found to be as effective as professional counselors on all criteria of counseling effectiveness; in fact, student counselors achieved significantly better results than did the professional counselors on the majority of variables used to measure the outcome of counseling. Furthermore, student counselors received a greater degree of acceptance from the counselees, and their counselees made significantly greater use of the information received during counseling as reflected by first semester grade point averages and residual study habits.

This review does not exhaust or should it limit consideration of the settings or situations in which college students can be used as therapeutic agents. Programs such as suicide prevention centers, groups for children of drug addicts or alcoholics, college sensitivity groups tutoring underprivileged children, teaching delinquent girls good grooming, visitations to the geriatric homes, and recreation therapy in schools for retarded children are all seeking motivated, sensitive individuals with innovative approaches who can fill their manpower need.

PROBLEMS OF USING STUDENT VOLUNTEERS

Not all of the literature concerning the use of student volunteers in the mental health field is positive. Umbarger et al. (1962) discussed the resistance of the mental hospital's ward staff to the student volunteer and attributed much of the resistance to staff jealousy; that is, students have time and opportunity for personal contact with the patients, whereas staff is kept busy with maintenance. An article in the *Pennsylvania Psychiatric Quarterly* (Volunteer Coordinators Resources, 1966) suggested that resistance to the volunteers occurs because of three reasons: (*a*) lack of information on the part of ward personnel, (*b*) insecurity on the part of the staff, and (*c*) the image of the volunteer as a nuisance rather than an intricate part of the program.

Haun (1968) raised a strong objection to the establishment of "lay administered ghettos for chronically ill," but Rioch (1966) suggested a flexible attitude and suggested that psychologists and psychiatrists are not exempt from feelings of jealousy.

> If we have invested long years of hard work in achieving a high professional status, including many courses that were dull and many examinations that were nerve wracking, and we are told that some young bit of a girl with no training can do the job as well as or better than we can, it is natural that we should try to find some objections [Rioch, 1966, p. 291].

Reiff (1967) discussed the dangers of a power struggle between professionals and nonprofessionals and the tensions surrounding such struggles. Perhaps the power struggle has already begun as the problem of controls over nonprofessionals is considered by Blau (1969a). He suggested that professionals should begin now to ascertain kinds of control over nonprofessionals which can and should be established. He suggested that the controls would be not only to ensure protection of the service recipient but also to ensure the effectiveness of the service rendered.

Another practical problem concerning the use of college students as psychotherapeutic agents has to do with the kind and amount of training needed prior to student-patient contact. The question of training of nonprofessionals is a topic worthy of a separate review, and numerous experimenters have investigated the problem (Berenson, Carkhuff, & Myrus, 1966; Carkhuff, 1968; Felzer, Bullock, Kay, & Wolf, 1968; Hart, Kroll, Berkowitz, & Woolcock, 1968; Linden & Stollak, 1969; Parsons & Parker, 1968; Sines, Silver, & Lucero, 1961; Truax, 1967; Vesprani, 1969).

While their naiveté has previously been presented as a potential advantage of using college students, it also is possible that it will be a liability. The unsophisticated student could project his own difficulties onto his patients or burden the patient with his own personal problems. Without intending harm, the college student may be tempted to "play" at psychotherapy by asking personal questions and attempting to interpret the patient's verbalizations. An additional potential pitfall of using college students in the mental health setting is that they tend to be temporary participants who have a peak experience in a setting without future responsibility. Since they are not forced to realize or continue with the problems encountered by professionals, they may not feel the responsibility for the long-term welfare of the patient. A further possibility is that because they have nothing to lose, that is, professional status, they may exploit the position or relationship with the patient to satisfy their own needs.

There is also concern with the Hawthorne effect. There is no way of knowing to what degree the student's enthusiasm and dedication is a function of newness to the situation and/or professional attention. It is possible that the pessimistic institutional climate of the setting may serve to dampen the student's vigor after a few months. Not only the total efficacy of college students but their relative efficacy at the beginning and at varying periods of time should be investigated. If it is found then that the college student agents lose their enthusiasm after a certain period of time, the time each student is involved in a given program could be limited, thus giving the patients the advantage of only peak experiences of the student therapeutic agent. It is hoped, however, that the freshness and strengths of the college student will offset many of the inevitable difficulties created by their utilization. As Cowen (1967) reflected: "The mental health movement has probably suffered more from its errors of omission than from its errors of commission, and we would be well advised to begin to correct some of the former, even at risk of the latter [p. 431]."

SUMMARY

The most obvious findings of this review are that there are few studies dealing with college students as therapeutic agents, and, further, most of those investigations which have been conducted are so methodologically inadequate that it is impossible to draw firm conclusions about the relative effectiveness of college students as therapeutic agents.

The studies are plagued with inadequacies in design which are characteristic of psychological studies in general and of clinical studies in particular. But even more distressing is the fact that few of the investigations of efficacy of college students as therapeutic agents utilized even the basic necessities of scientific inquiry. For example, less than 25% boasted a control group of untreated counselees. Also, only 5 of these 19 studies used pre- and posttesting and only 5 of 19 used objective measures (Cowen, 1968; Cowen, Zax, & Laird, 1966; Poser, 1966; Zunker & Brown, 1966).

Another problem encountered in reviewing the literature in which college students are used as therapeutic agents is that few of the studies are similar enough in any respect to warrant conclusions in a given area. The populations of the studies are extremely diverse; for instance, Poser (1966) used chronic schizophrenics, Umbarger et al. (1962) used "chronic psychotics." Spoerl's (1968) population included hospitalized college students while Goodman's (1967) students worked with troubled boys. Stollak's (1969) play therapists worked with children of unstated diagnoses from Michigan State Psychological Clinic. Also among the studies there is little consistency as to the kind or the amount of training. College students in some studies were given no training (Ralph, 1968; Spoerl, 1968), while others were given very specific training (Linden & Stollak, 1969; Stollak, 1969; Zunker & Brown, 1966). Further, there was a great deal of difference in motivation among the students. Some students received monetary remuneration (Goodman, 1967; Poser, 1966; Scheibe, 1965); others received college credit (Umbarger et al., 1962); and still others received no extrinsic remuneration (Levine, 1966; Spoerl, 1968). There were also differences in the duration and frequency of students working. Some students worked one day a week (Spoerl, 1968; Umbarger et al., 1962) while others worked full time (Lawton & Lipton, 1963; Poser, 1966).

Although there are not enough data from well-controlled studies to warrant conclusions concerning the relative efficacy of college students as therapeutic agents to patients, there is sufficient evidence to conclude that the therapeutic relationship has a definite, positive effect upon the college student therapist. The studies supporting this conclusion are much more methodologically complete than those describing the effect of the student therapist on patients; for instance, more than 60% of the former studies utilized control groups of students who did not participate in mental health programs. Further, over 90% of the studies describing the effects upon student therapists as a result of working in mental health settings used pre- and posttests which were relatively objective.

Personality changes such as positive changes in self-acceptance and moral judgments of a sexual and aggressive nature, greater self-understanding, increased self-confidence, and enhanced identity formation have been noted by experimenters in college students participating in mental health programs. It is felt that these studies are valid and that working in mental health programs may foster personality development in students in college. It is expected that college students are not unique in manifesting these changes, but most studies of nonprofessionals in mental health, other than college students, have not focused on these personality changes. One explanation is that personality changes in noncollege student volunteers participating in a mental health program occur less frequently or are less intense than changes in student volunteers. It may be that college students are more open to change, since goals and interests have not yet crystallized.

Attempts to conceptualize and explain the personality changes which are not unlike those reported in psychotherapy have been undertaken by Holzberg, Knapp, and Turner (1966). They have presented four formulations. Their first formulation views the personality changes in the students as a function of a heightened sense of self-competence, while a second hypothesis suggests that changes occur as a function of an enlarged and more flexible system of personality constructs. The clarification of their self-concepts and identities is also seen as a basis for personality change. The fourth formulation suggests an "emotional catharsis." One of the interpretations presented of how catharsis takes place is that "certain emotions may be extinguished by saturation, just as certain persistent habits may be overcome through excessive practice [p. 404]."

Reiff and Riessman (1965) hypothesized that added responsibility for a meaningful job, as well as the satisfaction garnered from acquiring new skills, may account for positive personality changes in some nonprofessionals. There also may be an increase in status and prestige as a result of working with professionals which is incorporated into their self-concept. Furthermore, one may gain added stability and security by reasoning that one must be in good shape himself in order to be able to help others.

A final hypothesis to explain positive personality changes occurring in nonprofessionals is an enhanced self-concept from the knowledge that he can influence the life of others. The fact that he can effect positive change in others may be particularly important to the college student who often feels like a computer card in a multi-university. It should be pointed out that these formulations are preliminary and have value only if they can be stated operationally and submitted to a direct test under controlled conditions.

It is felt that college student volunteer programs in mental health serve a variety of functions, including the providing of meaningful educational experiences for college students, alleviating the manpower squeeze in mental health, enhancing student development programs, preparing future generations of American leaders for constructive roles in mental health, and last, but particularly important for the present, is the providing of students with an opportunity to effect real and meaningful changes in their environment—a constructive cause to which they can devote their energies.

A number of problems face college student volunteer programs in mental health. Acceptance by staff and even by professionals has improved, but resistance to college students acting as therapeutic agents still exists. Another very closely related problem is the definition of interaction appropriate to nonprofessional college student therapeutic agents. Still a third problem area for the college student mental health program is the question of the kind and amount of training. Most agree that some training is necessary even if it is in the form of weekly supervision meetings. Diversity of opinion abounds, however, on the question of the exact program of training. Further, there is the question of the kinds of controls to be established over the nonprofessionals—organizational and legal controls as well as supervisory controls.

The use of nonprofessionals in general and college students in particular offer promise in the effort to provide more complete mental health services to an ever increasing population. However, researchers must move quickly with sound research designs to establish a firm empirical basis for justification of the nonprofessional as well as for student development in working with distressed individuals.

REFERENCES

Albee, G. W. *Mental health manpower trends.* New York: Basic Books, 1959.

Albee, G. W. The relation of conceptual models to manpower needs. In E. L. Cowen, E. A. Gardner, & M. Zax (Eds.), *Emergent approaches to mental health problems.* New York: Appleton-Century-Crofts, 1967.

Albee, G. W. Conceptual models and manpower requirements in psychology. *American Psychologist,* 1968, **23**, 317–320.

Arnhoff, F. N., & Jenkins, J. W. Subdoctoral education in psychology: A study of issues and attitudes. *American Psychologist,* 1969, **24**, 430–443.

Arnhoff, F. N., Rubinstein, E. A., & Speisman, J. G. *Manpower for mental health.* Chicago: Aldine, 1969.

Beck, J. C., Kantor, D., & Gelineau, V. A. Follow-up study of chronic psychotic patients "treated" by college case-aide volunteers. *American Journal of Psychiatry,* 1963, **120**, 269–271.

Belz, J. F., Drehmel, V. W., & Sivertsen, A. D. Volunteer: The community's participant in treatment of schizophrenic children in a day care program. *American Journal of Orthopsychiatry,* 1967, **37**, 221–222.

Berenson, B. G., Carkhuff, R. R., & Myrus, P. The interpersonal functioning and training of college students. *Journal of Counseling Psychology,* 1966, **13**, 441–446.

Blau, T. H. The professional in the community views the nonprofessional helper. *Professional Psychology,* 1969, **1**, 25–31. (a)

Blau, T. H. Psychologist views the helper. In C. Grosser, W. E. Henry, & J. G. Kelly (Eds.), *Nonprofessionals in the human services.* San Francisco: Jossey-Bass, 1969. (b)

Brennan, E. C. College students and mental health programs for children. *American Journal of Public Health,* 1967, **57**, 1767–1771.

Carkhuff, R. R. Lay mental health counseling: Prospects and problems. *Journal of Individual Psychology,* 1968, **24**, 88–93.

Christmas, J. J. Group methods in training and practice: Nonprofessional mental health personnel in a deprived community. *American Journal of Orthopsychiatry,* 1966, **36**, 410–419.

Cowen, E. L. Emergent approaches to mental health problems: An overview and directions for future work. In E. L. Cowen, E. A. Gardner, & M. Zax (Eds.), *Emergent approaches to mental health problems.* New York: Appleton-Century-Crofts, 1967.

Cowen, E. L. Effectiveness of secondary prevention programs using nonprofessionals in the school setting. *Proceedings of the 76th Annual Convention of the American Psychological Association,* 1968, **3**, 705–706.

Cowen, E. L. Combined graduate-undergraduate training in community mental health. *Professional Psychology,* 1969, **1**, 72–73. (a)

Cowen, E. L. Mothers in the classroom. *Psychology Today,* 1969, 3(7), 36–39. (b)

Cowen, E. L., Carlisle, R. L., & Kaufman, G. Evaluation of a college student volunteer program with primary graders experiencing school adjustment problems. *Psychology in the Schools,* 1969, **6**, 371–375.

Cowen, E. L., Izzo, L. D., Miles, H., Telschow, E. F., Trost, M. A., & Zax, M. A preventive mental health program in the school setting: Description and evaluation. *Journal of Psychology,* 1963, **56**, 307–356.

Cowen, E. L., Leibowitz, G., & Leibowitz, E. The utilization of retired people as mental health aides in the schools. *American Journal of Orthopsychiatry,* 1968, **38**, 900–910.

Cowen, E. L., Zax, M., Izzo, L. D., & Trost, M. A. Prevention of emotional disorders in the school setting: A further investigation. *Journal of Consulting Psychology,* 1966, **30**, 381–387.

Cowen, E. L., Zax, M., & Laird, J. D. A college student volunteer program in the elementary school setting. *Community Mental Health Journal,* 1966, **2**, 319–328.

Davison, G. C. The training of undergraduates as social reinforcers for autistic children. In L. Ullmann & L. Krasner (Eds.), *Case studies in behavior modification.* New York: Holt, Rinehart & Winston, 1966.

Department of Health, Education, and Welfare. Welfare Administration: Antidelinquency project measures effectiveness of volunteers. *Welfare in Review,* **4**, March 1966.

Donahue, G. T., & Nichtern, S. *Teaching the troubled child.* New York: Free Press, 1965.

Erikson, E. H. Identity and the life cycle: Selected papers. *Psychological Issues,* 1959, **1**, 1–171.

Fellows, L., & Wolpin, M. High school psychology trainees in a mental hospital. In B. G. Guerney, Jr. (Ed.), *Psychotherapeutic agents: New roles for nonprofessionals, parents, and teachers.* New York: Holt, Rinehart & Winston, 1969.

Felzer, S. B., Bullock, S., Kay, D., & Wolf, A. Mental health training program for community workers. *Proceedings of the 76th Annual Convention of the American Psychological Association,* 1968, **3**, 711–712.

Goodman, G. An experiment with companionship therapy: College students and troubled boys—assumptions, selection, and design. *American Journal of Public Health,* 1967, **57**, 1772–1777.

Gorlich, E. H. Volunteers in institutions for delinquents. *Children,* 1967, **14**, 147–150.

Greenblatt, M., & Kantor, D. An introduction. In C. Umbarger, J. Dalsimer, A. Morrison, & P. Breggin (Eds.), *College students in a mental hospital.* New York: Grune & Stratton, 1962. (a)

Greenblatt, M., & Kantor, D. Student volunteer movement and the manpower shortage. *American Journal of Psychiatry,* 1962, **118**, 809–814. (b)

Grosser, C., Henry, W. E., & Kelly, J. G. (Eds.) *Nonprofessionals in the human services.* San Francisco: Jossey-Bass, 1969.

Hart, N. J., Kroll, M., Berkowitz, D., & Woolcock, J. Assessment of demonstration community mental health program by its nonprofessional workers. *Proceedings of the 76th Annual Convention of the American Psychological Association,* 1968, **3**, 713–714.

Hartog, T. Nonprofessionals as mental health consultants. *Hospital and Community Psychiatry,* 1967, **18**, 223–225.

Haun, P. The future role of state mental institutions. *Psychiatric Quarterly,* 1968, **42**, 1–16.

Hersch, P. D., Kulik, J. A., & Scheibe, K. E. Personal characteristics of college volunteers in mental hospitals. *Journal of Consulting and Clinical Psychology,* 1969, **33**, 30–34.

Hirsch, I. S. Training mothers in groups as reinforcement therapists for their own children. *Dissertation Abstracts,* 1968, **28**, 4756B.

Hoch, E. L., Ross, A. O., & Winder, C. L. Conference on the professional preparation of clinical psychologists: A summary. *American Psychologist,* 1966, **21**, 42–51.

Holand, M. W., & Voss, F. H. Nontraditional assignments for volunteers. *Hospital and Community Psychiatry,* 1968, **19**, 221.

Holzberg, J. D. The companion program: Implementing the manpower recommendations of the joint commission on mental illness and health. *American Psychologist,* 1963, **18**, 224–226.

Holzberg, J. D., & Gewirtz, H. A method of altering attitudes toward mental illness. *Psychiatric Quarterly Supplement,* 1963, **37**, 56–61.

Holzberg, J. D., Gewirtz, H., & Ebner, E. Changes in moral judgment and self-acceptance as a function of companionship with hospitalized mental patients. *Journal of Consulting Psychology,* 1964, **28**, 299–303.

Holzberg, J. D., & Knapp, R. H. The social interaction of college students and chronically ill patients. *American Journal of Orthopsychiatry,* 1965, **35**, 487–492.

Holzberg, J. D., Knapp, R. H., & Turner, J. L. Companionship with the mentally ill: Effects on the personalities of college student volunteers. *Psychiatry,* 1966, **29**, 395–405.

Holzberg, J. D., Knapp, R. H., & Turner, J. L. College students as companions to the mentally ill. In E. L. Cowen, E. A. Gardner, & M. Zax (Eds.), *Emergent approaches to mental health problems.* New York: Appleton-Century-Crofts, 1967.

Holzberg, J. D., Whiting, H. S., & Lowy, D. G. Chronic patients and a college companion program. *Mental Hospitals,* 1964, **15**, 152–158.

Hunt, J. McV. Message, potpourri and the Hall-Nebraska "model." *Clinical Psychologist,* 1969, **22**, 127–136.

Johnston, R. Some casework aspects of using foster grandparents for emotionally disturbed children. *Children,* 1967, **14**, 46–52.

Kantor, D. Inducing preferences for mental health careers. *National Association of Social Workers,* 1959, 75.

Kantor, D. Impact of college students on chronic mental patients and on the organization of the mental hospital. *Proceedings of the College Student Companion Program Conference.* Stratford, Conn.: Connecticut State Department of Mental Health, 1962.

Keniston, K. College students and children in developmental institutions. *Children,* 1967, **14**, 2–7.

Knapp, R. H., & Holzberg, J. D. Characteristics of college students volunteering for service to mental patients. *Journal of Consulting Psychology,* 1964, **28**, 82–85.

Knott, P. D. On the manpower problem and graduate training in clinical psychology. *American Psychologist,* 1969, **24**, 675–679.

Kreitzer, S. F. College students in a behavior therapy program with hospitalized emotionally disturbed children. In B. G. Guerney, Jr. (Ed.), *Psychotherapeutic agents: New roles for nonprofessionals, parents, and teachers.* New York: Holt, Rinehart & Winston, 1969.

Lawton, M. P., & Lipton, M. B. Student-employees become companions to patients. *Mental Hospitals,* 1963, **14**, 550–556.

Levine, C. Impact of work with mental patients on student volunteers. *Journal of Human Relations,* 1966, **14**, 422–433.

Linden, J. I., & Stollak, G. E. The training of undergraduates in play techniques. *Journal of Clinical Psychology,* 1969, **25**, 213–218.

Madison, P. *Personality development in college.* Reading, Mass.: Addison-Wesley, 1969.

Mitchell, W. E. Amicatherapy: Theoretical perspectives and an example of practice. *Community Mental Health Journal,* 1966, **2**, 307–314.

Nichtern, S., Donahue, G. T., O'Shea, J., Marans, M., Curtis, M., & Brody, C. A community educational program for the emotionally disturbed child. *American Journal of Orthopsychiatry,* 1964, **34**, 705–713.

Parsons, L. B., & Parker, G. V. Personal attitudes, clinical appraisals, and verbal behavior of trained and untrained therapists. *Journal of Consulting Psychology,* 1968, **32**, 64–71.

Patterson, G. R., & Brodsky, G. A behavior modification programme for a child with multiple problem behaviors. *Journal of Child Psychology,* 1966, **7**, 277–295.

Pearl, A., & Riessman, F. (Eds.) *New careers for the poor: The nonprofessional in human service.* New York: Free Press, 1965.

Perlmutter, F., & Durham, D. Using teen-agers to supplement casework service. *Social Work,* 1965, **10**, 41–48.

Poser, E. The effect of therapists' training on group therapeutic outcome. *Journal of Consulting Psychology,* 1966, **30**, 283–289.

Ralph, D. E. Attitudes toward metal illness among two groups of college students in neuropsychiatric hospital setting. *Journal of Consulting and Clinical Psychology,* 1968, **32**, 98.

Reding, G. R., & Goldsmith, E. F. The nonprofessional hospital volunteer as a member of the psychiatric consultation team. *Community Mental Health Journal,* 1967, **3**, 267–272.

Reiff, R. Mental health manpower and institutional change. In E. L. Cowen, E. A. Gardner, & M. Zax (Eds.), *Emergent approaches to mental health problems.* New York: Appleton-Century-Crofts, 1967.

Reiff, R., & Riessman, F. The indigenous nonprofessional: A strategy of change in community action and community mental health programs. *Community Mental Health Journal,* 1965, Monograph No. 1.

Reinherz, H. Group leadership of student volunteers. *Mental Hospitals,* 1962, **13**, 600–603.

Reinherz, H. The therapeutic use of student volunteers. *Children,* 1964, **2**, 137–142.

Rioch, M. J. Changing concepts in the training of therapists. *Journal of Consulting Psychology,* 1966, **30**, 290–292.

Rioch, M. J., Elkes, C., Flint, A. A., Usdansky, B. S., Newman, R. G., & Silber, E. National Institute of Mental Health pilot study in training mental health counselors. *American Journal of Orthopsychiatry,* 1963, **33**, 678–689.

Rogers, C. R. The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology,* 1957, **21**, 95–103.

Rosenbaum, M. Some comments on the use of untrained therapists. *Journal of Consulting Psychology,* 1966, **30**, 292–294.

Sanford, N. Higher education as a social problem. In N. Sanford (Ed.), *The American college.* New York: Wiley, 1962.

Savino, M. T., & Schlamp, F. T. The use of nonprofessional rehabilitation aides in decreasing rehospitalization. *California Mental Health Research Digest,* 1967, **5**, 254–256.

Scheibe, K. E. College students spend eight weeks in mental hospital: A case report. *Psychotherapy: Theory, Research and Practice,* 1965, **2**, 117–120.

Sines, L. I., Silver, R. J., & Lucero, R. J. The effect of therapeutic intervention by untrained "therapists." *Journal of Clinical Psychology,* 1961, **17**, 394–396.

Spoerl, O. H. An activity-centered volunteer program for university students. *Hospital and Community Psychiatry,* 1968, **19**, 114–116.

Stollak, G. E. The experimental effects of training college students as play therapists. In B. G. Guerney, Jr. (Ed.), *Psychotherapeutic agents: New roles for nonprofessionals, parents, and teachers.* New York: Holt, Rinehart & Winston, 1969.

Stover, L., & Guerney, Jr., B. G. The efficacy of training procedures for mothers in filial therapy. *Psychotherapy: Theory, Research and Practice,* 1967, **4**, 110–115.

Truax, C. B. The training of nonprofessional personnel in therapeutic interpersonal relationships. *American Journal of Public Health,* 1967, **57**, 1778–1791.

Truax, C. B., & Wargo, D. G. Psychotherapeutic encounters that change behavior: For better or for worse. *American Journal of Psychotherapy,* 1966, **20**, 499–520.

Tuckman, J., & Lavell, M. Attrition in psychiatric clinics for children. *Public Health Report,* 1959, **74**, 309–315.

Umbarger, C. C., Dalsimer, J. S., Morrison, A. P., & Breggin, P. R. *College students in a mental hospital.* New York: Grune & Stratton, 1962.

Vesprani, G. J. Personality correlates of accurate empathy in a college companion program. *Journal of Consulting and Clinical Psychology,* 1969, **33**, 722–727.

Volkman, R., & Cressey, D. R. Differential association and the rehabilitation of drug addicts. In F. Riessman, J. Cohen, & A. Pearl (Eds.), *Mental health of the poor.* New York: Free Press, 1964.

Volunteer Coordinators Resources. Direction of change: Facets of volunteer resources. *Pennsylvania Psychiatric Quarterly,* 1966, **6**, 48–58.

Wahler, R. G., Winkel, G. H., Peterson, R. F., & Morrison, D. C. Mothers as behavior therapists for their own children. *Behavior Research and Therapy,* 1965, **3**, 113–124.

Walker, C. E., Wolpin, M., & Fellows, L. The use of high school and college students as therapists and researchers in a state mental hospital. *Psychotherapy: Theory, Research and Practice,* 1967, **4**, 186–188.

Yablonsky, L., & Dederich, C. E. Synanon: As a program for training ex-offenders as therapeutic agents. In, *Experiment in culture expansion.* Sacramento, Calif.: California Department of Corrections, 1963.

Zax, M., & Cowen, E. L. Early identification and prevention of emotional disturbance in a public school. In E. L. Cowen, E. A. Gardner, & M. Zax (Eds.), *Emergent approaches to mental health problems.* New York: Appleton-Century-Crofts, 1967.

Zimmerman, E. H., & Zimmerman, J. The alteration of behavior in a special classroom situation. *Journal of the Experimental Analysis of Behavior,* 1962, **5**, 59–60.

Zunker, V. G., & Brown, W. F. Comparative effectiveness of student and professional counselors. *Personnel and Guidance Journal,* 1966, **44**, 738–743.

(Received February 26, 1970)

Psychological Bulletin
1971, Vol. 76, No. 2, 128–148

VALIDATION AND EXTENSION OF THE CONTINGENCY MODEL OF LEADERSHIP EFFECTIVENESS: A REVIEW OF EMPIRICAL FINDINGS[1]

FRED E. FIEDLER[2]

University of Washington

This paper reviews studies which tested and extended the contingency model of leadership effectiveness. The model predicted a curvilinear relationship such that leaders with low least preferred co-worker (LPC) scores ("task-oriented") would perform more effectively in very favorable and unfavorable situations, while high LPC leaders ("relationship-oriented") would perform more effectively in situations intermediate in favorableness. The model is supported by data from field studies but not fully supported by data from laboratory studies. Field and laboratory studies extending the model supported the prediction that the situational favorableness dimension moderates the relationship between leadership style and group or organizational performance. Relations between leadership style and group performance for the coacting task groups were similar to those for interacting groups within the same group classification of the model, while training groups appear to require a predominantly relationship-oriented leadership style, irrespective of the type of leadership situation.

A contingency model of leadership effectiveness, described in a theoretical paper 7 years ago (Fiedler, 1964), has stimulated numerous studies in the area testing the model as well as attacking it in a recent issue of this journal (Graen, Alvares, Orris, & Martella, 1970). The present paper reviews 25 investigations purporting to test or extend the model.

The contingency model postulates that the performance of interacting groups is contingent upon the interaction of leadership style and situational favorableness. It has been suspected for some time that group effectiveness depends on attributes of the leader as well as of the situation (e.g., Tannenbaum & Schmidt, 1958; Terman, 1904). The question in leadership theory has been, What kind of leadership style for what kind of situation? The contingency model specifies that the so-called "task-oriented" leaders perform more effectively in very favorable and very unfavorable situations, while "relationship-oriented" leaders perform more effectively in situations intermediate in favorableness. The theory operationalizes leadership style as well as situational favorableness and, therefore, lends itself to empirical testing.

This study first defines the main terms of the theory, and briefly reviews the findings on which the model is based. It then presents (*a*) validation evidence relevant to the model's prediction of group performance in real-life studies and laboratory experiments; (*b*) extensions of the model to a more broadly defined hypothesis; and (*c*) an analysis of results bearing upon the reclassification and prediction of performance of coacting groups.

Definitions

The main terms of the theory—*leadership style*, *situational favorableness*, and *leadership or group effectiveness*—are briefly described below, as are the definitions of *interacting* and *coacting* groups.

Interacting groups. These are groups in which the members work cooperatively and interdependently on a common task. The contributions of individual members of these groups cannot, therefore, readily be isolated, and the members, for this reason, are typically re-

---

[1] This review was prepared under Contract N00014-67-A-0103-0012 with the Office of Naval Research and the University of Washington, Seattle (Fred E. Fiedler, Principal Investigator). Research was supported in part by Contract N00014-67-A-0103-0013 with the Advanced Research Projects Agency of the Office of Naval Research.

[2] Requests for reprints should be sent to Fred E. Fiedler, Department of Psychology, University of Washington, Seattle, Washington 98105.

I am indebted to my colleagues, Anthony Biglan, Uriel Foa, Terence R. Mitchell, and Gerald Oncken for their invaluable criticisms and suggestions for the successive manuscripts.

warded or penalized as a group. In contrast, in *coacting* groups members perform their tasks in relative independence of one another, as for example, members of bowling teams, men in piecework production or in training situations in which each participant typically receives an individual score or evaluation at the end of training.

Leadership style. The predictor measure used in studies of the contingency model is the least preferred co-worker (LPC) score. This score is obtained by first asking an individual to think of all co-workers he has ever had. He is then asked to describe the one person with whom he has been least able to work well, that is, the person he least prefers as a co-worker. This need not be someone with whom he works at the time. The description is made on 8-point, bipolar adjective scales, for instance,

friendly :——:——:——:——:——:——:——:——: unfriendly
8 7 6 5 4 3 2 1

cooperative :——:——:——:——:——:——:——:——: uncooperative
8 7 6 5 4 3 2 1

As a rule, 16 to 24 items have been used in LPC scales. The LPC score is obtained by summing the item values, giving a value of 8 to the favorable pole of each scale. Thus, a high score indicates that the subject has described his least preferred co-worker in relatively favorable terms, that is, with an average item value in the neighborhood of 5 on the 8-point scale. A low score means that the least preferred co-worker is described in a very negative, rejecting manner, that is, an LPC score of about 2 (Fiedler, 1967a, p. 43). It should also be noted that the low LPC person describes his least preferred co-worker in a uniformly, hence undifferentiated or stereotyped manner as "all bad." The high LPC person's description has a considerably greater item variance (a standard deviation of 1.43 for the high versus .43 for the low LPC person).

The score has been difficult to interpret. While labels of relationship-oriented versus task-oriented have been given to high versus low LPC persons, the terms are somewhat misleading. First, only in situations which are unfavorable (that is, stressful, anxiety arousing, giving the leader little control) do we find leader behaviors which correspond to these terms (Fiedler, 1967a). Second, Mitchell (1970) has found evidence that high LPC leaders tend to be cognitively more complex in their thinking about groups, while low LPC leaders tend to give more stereotyped cognitively simple responses. Similar results (i.e., a correlation of .35) have been reported by Schroder and his co-workers (H. Schroder, personal communication, 1969).

Thus, the LPC score must be seen as a measure which at least in part reflects the cognitive complexity of the individual and which in part reflects the motivational system that evokes relationship-oriented and task-oriented behaviors from high versus low LPC persons in situations which are unfavorable for them as leaders.

Situational favorableness. The variable that moderates the relationship between LPC and group performance is the situational favorableness dimension. It is conceptually defined as the degree to which the situation itself provides the leader with potential power and influence over the group's behavior. Situational favorableness appears to be quite important in affecting a wide range of group phenomena, as well as interpersonal behaviors. It seems likely that this dimension may have far-reaching significance in other personality research, as well as in social psychological investigations.

Situational favorableness has been operationalized in a number of ways which are discussed later. The original work on the contingency model presented one method based on three component dimensions that affect the degree to which the situation provides the leader with potential power and influence. These are leader-member relations, task structure, and position power. The hypothesis was that (*a*) it is "easier" to be a leader of a group that respects and accepts its leader, or in which

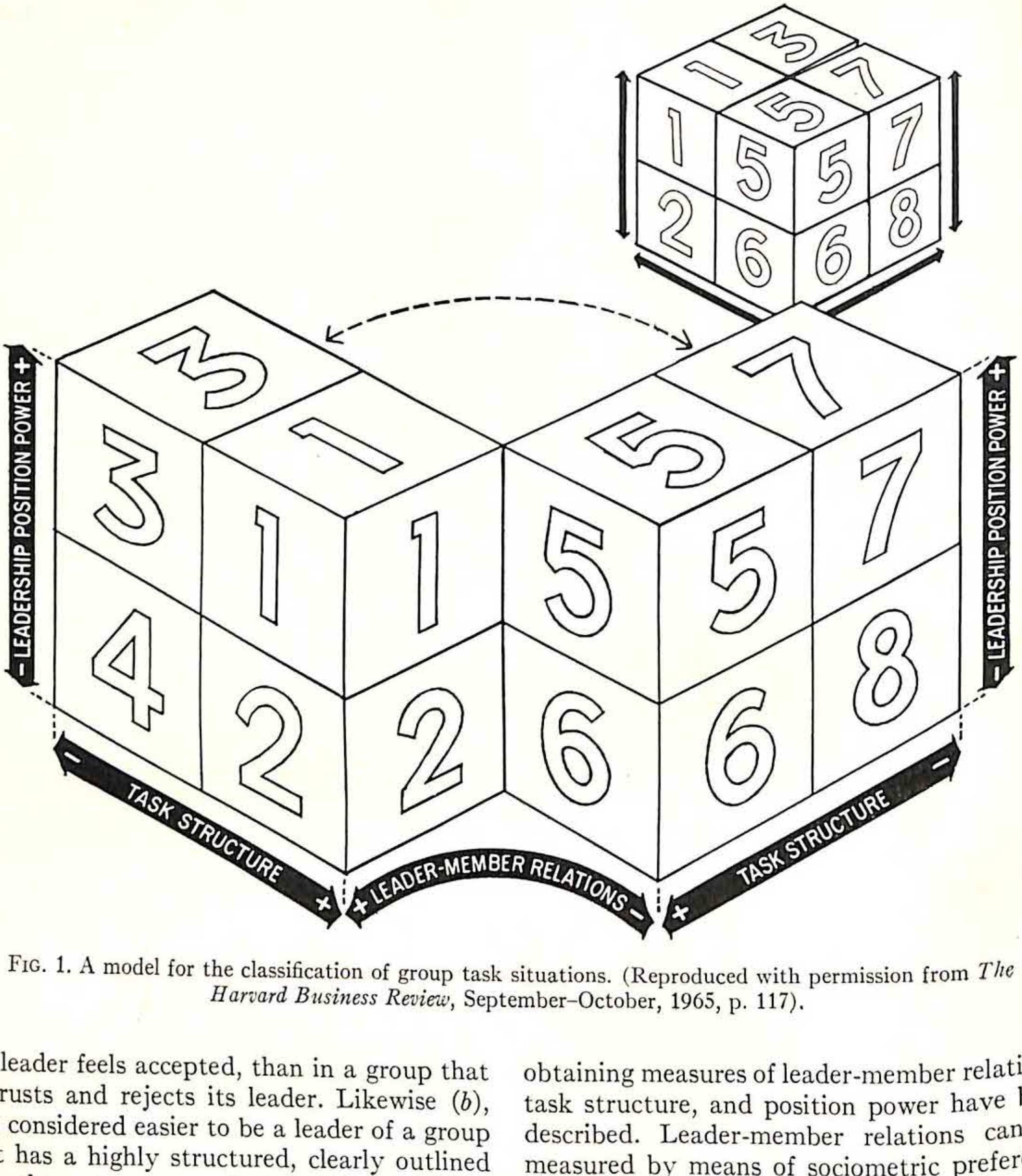

3
1
5
7
1
5
5
7
2
6
6
8
3
1
5
7
3
1
1
5
5
7
4
2
2
6
6
8
– LEADERSHIP POSITION POWER +
– TASK STRUCTURE +
+ LEADER-MEMBER RELATIONS –
+ TASK STRUCTURE –
– LEADERSHIP POSITION POWER +

FIG. 1. A model for the classification of group task situations. (Reproduced with permission from *The Harvard Business Review*, September–October, 1965, p. 117).

the leader feels accepted, than in a group that distrusts and rejects its leader. Likewise (*b*), it is considered easier to be a leader of a group that has a highly structured, clearly outlined task than of a group that has a vague, unstructured, nebulous task; (*c*) it is easier to be a leader when the position is vested with power (when the leader has the power to hire and fire, promote and transfer, give raises or lower salaries) than it is to be a leader who enjoys little or no power over his members: It is easier to be a general manager than the chairman of a volunteer group.

Leader-member relations were considered to be the most important of these situational factors, and subsequent studies have supported this supposition (Fishbein, Landy, & Hatch, 1969; Mitchell, 1969). Detailed instructions for obtaining measures of leader-member relations, task structure, and position power have been described. Leader-member relations can be measured by means of sociometric preference ratings or by a group atmosphere scale which is similar in form and content to LPC, but asks the subject to rate his group as a whole. Scales for rating task structure and position power are described in Fiedler (1967a, pp. 24, 28, 269, 281–291).

We could then classify group situations by means of the three dimensions. For this purpose, all groups were classified as falling above or below the median on each dimension. This led to an eight-celled classification system which can be depicted as an eight-celled cube (Figure 1). Each of the eight cells or "octants" can be scaled in terms of how much power and

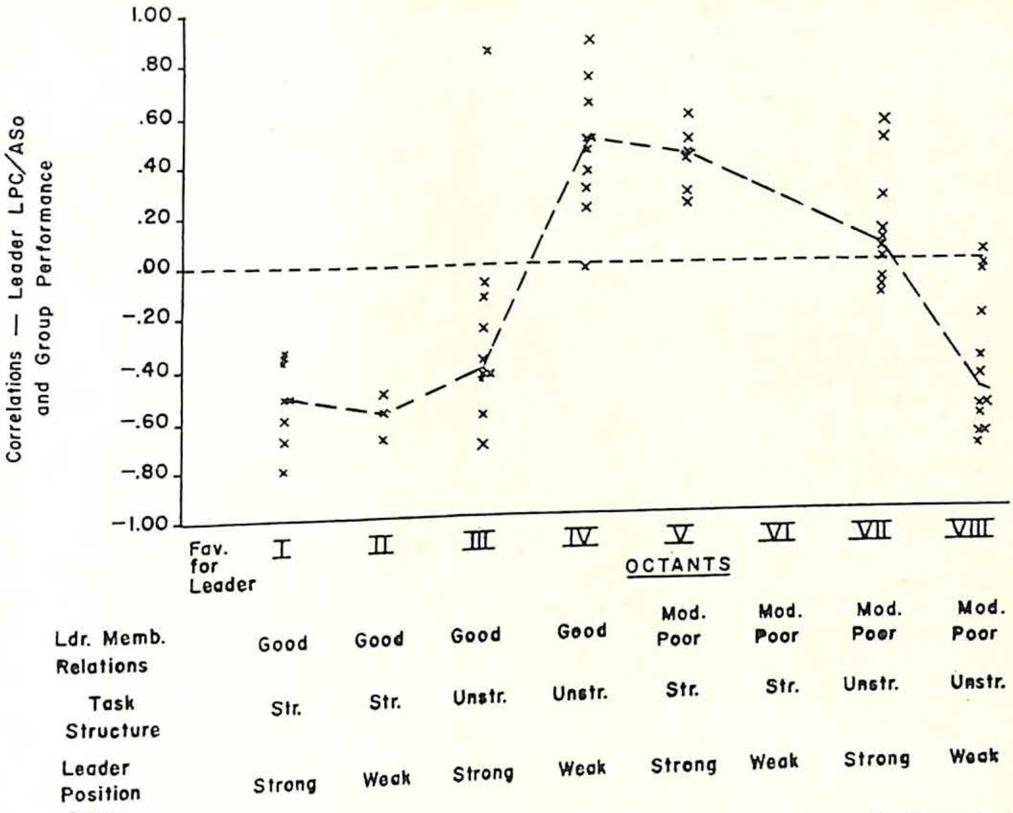


FIG. 2. Correlations between leadership, LPC scales, and group effectiveness plotted for each cell or octant of the situational favorableness dimension for studies of interacting groups conducted prior to 1963.

influence a leader might have in such a situation. Obviously, a liked and accepted leader who has a clear-cut task as well as power over the fate of his members (Octant I) will have a very favorable situation. Conversely, a distrusted chairman of a volunteer group with a vague problem-solving or policy-making task (Octant VIII) will be in a very unfavorable situation to exert power. Other octants fall between these two extremes.

Leadership effectiveness. The performance of the leader is here defined in terms of the major assignment of the group; that is, the leader's effectiveness is measured on the basis of the group's performance of its major assigned task. While such other aspects of group behavior as morale, member satisfaction, or personal growth might be important concomitants of group effectiveness, they are here not considered to be the primary criterion, but rather contributors to performance. In other words, we evaluate the performance of an orchestra conductor not by his ability as a musicologist or the happiness of his musicians, but by how well his orchestra plays. Whether happy musicians play better than unhappy musicians, or whether the man who is a great musicologist is a better conductor is an important research question in its own right. The major question asked here is the relationship of leadership style (specifically LPC) and group or organizational effectiveness.

Previous Results

Interacting groups from 15 studies, antedating 1963, were classified according to their situational favorableness, and the correlation between leader LPC and performance was then computed for each set of groups. The correlations between the leader's LPC score and the group's effectiveness measures, when plotted against situational favorableness, generated a bow-shaped distribution indicating that the low LPC leaders performed more effectively than high LPC leaders in very favorable and very unfavorable situations; high LPC leaders performed more effectively in situations intermediate in favorableness (see Figure 2).

We here review the validation evidence that has accumulated since publication of the model in 1964. Before specifically reporting any of the studies, it should be stressed that the group classification system was viewed "as a very convenient starting point for presenting the empirical results which we have obtained in our research on interacting groups [Fiedler, 1967a, p. 34]." Improved methods for measuring situational favorableness were expected to be developed in time. The three component dimensions did, however, turn out to be a very convenient method for testing the model since operationalized measures were available.

VALIDATION EVIDENCE OF THE CONTINGENCY MODEL

Classification of Studies

A number of studies designed to test the model have been conducted by various independent investigators as well as by the writer and his associates. Some of these investigators, by design, and others, by oversight, have not followed the methodology originally described. Different operationalizations of situational favorableness were used in some studies, while others extended the model to coacting groups (e.g., Hill, 1969; Hunt, 1967), and some used leadership style measures unrelated to LPC (Shaw & Blum, 1966). These differences in methodology and divergencies from the model are, of course, quite appropriate and desirable. However, studies that do not conform to the explicit methodology of the earlier work cannot be used as exact tests of the model.

This divergence in method presents difficulties only where the investigator and this reviewer disagree on the appropriateness of a study for testing the contingency model, or where the methodology is inadequate to test the model. This problem has here been handled as follows:

1. Four independent judges carefully read the definition of interacting and coacting groups, and the definitions of the various subdimensions of the situational favorableness dimension presented by Fiedler (1967a).
2. The judges were then given all studies that purported to test the contingency model. They were asked to read the entire methodology section in the case of shorter investigations, and relevant sections of very extensive studies. They were not asked to read the results. Using the scales described by Fiedler (1967a), the judges classified each study in terms of the group situation into which it should be classified.
3. Groups were included among the validation studies of interacting groups if three of the four judges could agree that the groups were interacting, and if at least three of the judges agreed into which octant the groups should be classified.

Because of the nature of the pre-1963 research on which the original analysis was based, all groups in Octant I, II, and V were from field studies of natural groups and organizations, and all but one set of groups in Octants III, IV, VII, and VIII were from laboratory experiments using ad hoc groups. It is very difficult to reproduce certain effects commonly found in natural groups under experimental conditions. These include, for example, high leader position power, high stress, and very poor leader-member relations. For this reason, the model can be more meaningfully evaluated as a predictor if natural groups and ad hoc groups in laboratory experiments are considered separately, as well as together.

Field Studies

The Hunt studies. The first field study to test the contingency model was conducted by Hunt (1967) in three different organizations, namely, a large physical science research laboratory, a chain of supermarkets, and a heavy machinery plant. In each case Hunt obtained group atmosphere scores from managers and foremen, and ratings of position power and task structure from management personnel at higher levels in the hierarchy. Higher management of the research laboratory and in the manufacturing plant also provided performance ratings, while objective criteria based on an index of sales per employee manhour was derived for managers of meat markets.

The position power of all managers and foremen was judged to be high. Task structure was rated to be high for developmental research and for meat market managers, and low for managers of basic research groups and for general foremen of the heavy machinery plant. The correlations obtained for each of these groups, by octant, are presented in Table 1. (Hunt's results on coacting groups are presented in a later section.)

TABLE 1

TEST OF THE CONTINGENCY MODEL IN HUNT'S STUDIES

Sample	Octants			
	I	III	V	VII
Research chemists: basic research		.60 (6)		.30 (5)
Research chemists: development	−.67 (7)			
Meat markets	−.51 (10)		.21 (11)	
General foremen: heavy manufacturing		−.80 (5)		−.30 (5)

Note.—Numbers in parentheses indicate number of cases.

Electronics firm. A second set of real-life groups was investigated by Hill (1969). The research was conducted in a large electronics firm. The study dealt with supervisors of engineering teams and with instructors of assembly groups. Assembly-line instructors were rated by Hill's judges as having structured tasks and low position power, while supervisors of engineering groups were rated as having unstructured tasks and high position power. Leader-member relations were measured by supervisors' group atmosphere scores, trichotomized, with the upper third considered to have good leader-member relations, and the lower third considered to have poor leader-member relations.

The correlation for nine assembly instructors' teams with high group atmosphere (Octant II) was − .10, and for the nine teams with low group atmosphere (Octant IV), − .24. The correlation between eight engineering supervisors' LPC and performance for high group atmosphere (Octant III) was − .29, and for eight supervisors with low group atmosphere groups (Octant VII), it was .62.

Public Health Teams: I. Fiedler, O'Brien, and Ilgen (1969) conducted a study of public health volunteers in Honduras during the summer of 1966. The sample consisted of 225 teenagers who were assigned to teams in Honduras to operate public health clinics and to perform community development work in outlying towns and villages. Formal leaders were not assigned. The teams' informal leaders were identified at the end of the volunteer period on the basis of sociometric questionnaire responses. The task of these groups, namely, to run a public health clinic and, time permitting, to perform some community development work, was fairly well-specified by the sponsoring organization. Problems arose when the villagers and the town officials failed to cooperate and when the population was unsupportive or hostile. Under these conditions the volunteers experienced considerable stress, often verbalized as "feeling at a loss about what to do." Under these stressful conditions, the team members were more or less on their own in trying to cope with the problems they encountered. All judges who evaluated the study agreed that the position power of the leader was low, and three of four judges considered the task to be relatively structured in the stress-free condition, but unstructured in the situation in which village support was absent.

The informal leader's group atmosphere scores were used to measure the leader-member relations. Groups operating in cooperative or favorable villages could then be classified as falling into Octants II or VI, depending on the leader's group atmosphere score; groups in uncooperative, unfavorable villages could be classified as falling into Octants IV or VIII, again depending on the leader's group atmosphere scores.

Public Health Teams: II. A second study, practically identical in procedure, was conducted in the same organization (but with different volunteers) 2 years later by O'Brien, Fiedler, and Hewett.[3] One major difference in training was that most volunteers had received a culture training program developed for use of the project. This program made the situation somewhat more favorable since it provided some information that enabled the volunteers to understand and to communicate more effectively with their host nationals. A study of the

---

[3] O'Brien, G. E., Fiedler, F. E., & Hewett, T. The effects of programmed culture training upon the performance of volunteer medical teams in Central America. Urbana, Illinois: Group Effectiveness Research Laboratory, University of Illinois, 1969.

TABLE 2

Correlations between Leader LPC Scores and Rates: Team Performance in Two Studies (1966, 1968) of Public Health Volunteer Teams

Situation			Octant	1966		1968		Median
Group atmosphere	Task structure	Position power		Correlation	n	Correlation	n	
High	High	Low	II	−.21	13	−.46	7	−.33
High	Low	Low	IV	.00	15	.47	9	.23
Low	High	Low	VI	.67*	9	−.45	8	.11
Low	Low	Low	VIII	−.51	12	−.14	7	−.32

* $p < .05$.

effects of this program on adjustment and performance, conducted in 1967, provided evidence that individuals who had received "culture assimilator" training had adjusted and performed more effectively abroad than those who had not.

The correlations between leader LPC and performance ratings of the headquarters staffs are presented in Table 2 for both studies. The median correlations for these groups support the model (correlations for Octant VI were not predicted).

The patterns of the 1966 and 1968 correlations both form a bow-shaped pattern, but the positive correlation was found in Octant VI in the 1966 group, but in Octant IV in the 1968 group.

Unclassifiable Field Studies

A number of other studies designed to test the contingency model cannot be included in the present analysis because the judges rating the published research could not agree on the classification of the groups. Thus, a major study by Butterfield (1968) attempted to compare five theories of leadership, namely, four theories developed at the University of Michigan and the contingency model. The study was conducted in an administrative unit of a federal agency. The sample of groups was not clearly described. References to groups indicated, however, that "the various work units were engaged in rather different functions ranging from delivering mail to performing financial audits . . . [p. 61]," and measures of effectiveness such as "typed papers, delivered mail, dispatched automobiles, coding systems . . . [Appendix B]." These descriptions led all four judges to the conclusion that a number of groups must have been coacting (e.g., messenger services, typing pools, motor pool dispatchers, etc.). Equally important, however, Butterfield's measure of leader-member relations consisted of only two questionnaire items, one of which dealt with how much annoyance an employee feels with the manager. The important leader-member relations dimension was, therefore, not adequately represented in Butterfield's study.

A field study by Kretzschmar and Lueck (1969) dealt with 67 managers in business administrations of four industrial companies in Germany. In contrast to other studies, all measures of task structure, leader-member relations, and position power were based on supervisors' own ratings. In addition, supervisors also rated their own effectiveness. Our studies have shown that leaders' evaluations of their own performance typically do not correlate with objective measures, and these methodological differences made it impossible to compare this study with others in the group.

The field studies which can be used to test the contingency model are, then, those by Hunt (1967), by Hill (1969), by Fiedler et al. (1969), and by O'Brien et al. (see Footnote 3). These results are now discussed, along with those obtained in laboratory studies.

Laboratory and Field Experiments

The contingency model was based on field studies in Octants I, II, and V, and on controlled experiments in Octants III, IV, VII, and VIII. Since that time a number of field and laboratory experiments have attempted to test the predictions of the model in all octants.

The Belgian Navy Study. This was a large field experiment conducted in cooperation with

TABLE 3

CORRELATIONS BETWEEN LEADER LPC SCORES AND GROUP PERFORMANCE SCORES OBTAINED IN STRUCTURED AND UNSTRUCTURED TASKS OF THE BELGIAN NAVY STUDY

Homogeneous groups						Heterocultural groups					
Octant	GA	TS	PP	Correlations[a]		Octant	GA	TS	PP	Correlations[a]	
I	+	+	+	−.72	−.77	IX	+	+	+	.03	.77
II	+	+	−	.37	.50	X	+	+	−	.77	−.53
III	+	−	+	−.16	−.54	XI	+	−	+	.20	−.26
IV	+	−	−	.08	.13	XII	+	−	−	−.89	.70
V	−	+	+	.16	.03	XIII	−	+	+	.08	−.19
VI	−	+	−	.07	.14	XIV	−	+	−	.53	−.90
VII	−	−	+	.26	−.27	XV	−	−	−	−.37	.08
VIII	−	−	−	−.37	.60	XVI	−	−	−	−.36	−.60

Note.—GA = group atmosphere, TS = task structure, PP = position power.
[a] Two correlations were obtained per cell corresponding to the order of presentation ($n = 6$).

the Belgian naval forces. The study involved 96 three-man teams which were experimentally assembled so that 48 teams would be culturally homogeneous (all French or all Dutch speaking) and 48 teams were heterogeneous (the leader from one language sector and the members from the other). In 48 teams the leader was a petty officer who had high position power, while the other 48 teams were headed by a recruit who had low position power. The groups were given four different tasks, one of which was coacting (teaching men to assemble an automatic pistol). The coacting task and one of the two structured tasks could not be reliably scored, and only one structured and one unstructured task were, therefore, suitable for more intensive analysis.[4] Leader-member relations were assessed on the basis of group climate scores obtained from leaders after each of the task sessions.

Since the contingency model, presented in 1964, was based on culturally homogeneous groups, that is, groups in which all members had the same mother tongue, only these will be used for validation purposes at this time. Table 3 presents the correlations between leader LPC scores and group performance scores. Two correlation coefficients, each with $n = 6$, were computed since the tasks were presented in counterbalanced order. As can be seen, the findings do not support the contingency model which postulates a curvilinear relationship.

A post facto analysis of the data suggested that the laboratory manipulations had not been adequate for the purpose of creating a sufficiently unfavorable situation for the leader. The homogeneous groups probably did not develop really poor leader-member relations, and there was also a question whether the supposedly structured task (requiring the group to find the shortest route for a ship which had to cover 12 ports) was sufficiently structured. The task, basically a topological puzzle, requires problem solving, and subsequent studies have shown this type task to be intermediate in structure. A bow-shaped relationship similar to that of the contingency model did emerge when heterocultural groups were included in the analysis to intensify the situational difficulty for the leader. Heterocultural groups are, of course, much more difficult to handle, and leaders reported significantly higher anxiety and greater tension in heterogeneous groups than in homogeneous groups. This effect was, however, not predicted.

The Japanese student study. Shima (1968) tested the contingency model in Japan, using two Guilford[5] tests, namely, the Unusual Uses

---

[4] The Graen et al. review (1970) included all three interacting tasks in the analysis of evidential results of the contingency model. However, it was clearly pointed out by the present writer (1967a, p. 161) that the first structured task was methodologically inadequate: 9 groups obtained perfect scores and 62 groups made a total of 189 errors by not following instructions. The first task was, therefore, quite unsatisfactory and could not very well have yielded anything but random results.

[5] Guilford, J. P., Berger, R. M., & Christiansen, P. R. A factor analytic study of planning I: Hypothesis and description of tests. Los Angeles, California: Psychological Laboratory, University of Southern California, 1954.

Test considered to be moderately structured, and an "integration" task which required groups to invent a story using 10 unrelated words. All subjects were high school students, and these were assembled into 32 groups. The leaders were elected by the group members, and the group's relations with the leader were, therefore, assumed to be good. However, the judges rating this study disagreed with Shima's classification of position power. Since the leader was elected and, therefore, had obtained his position from his fellow students, and since the groups were ad hoc, the judges rated position power in this study as being low. Based on the assumption that leader position power was low, the groups would be classified as falling into Octants II and IV. The corresponding correlations were $-.26$ ($n = 16$) and .71 ($n = 16$, $p < .05$), thus supporting the model.

Church leadership groups. Mitchell (1969) conducted a small study of group performance as part of a leadership training workshop to determine the relationship between cognitive complexity and LPC. The participants were members of Unitarian churches who attended a leadership workshop. Each of the groups performed two of four tasks, one structured, and one unstructured. The tasks consisted of finding the shortest route for (*a*) a school bus and (*b*) a cross-country road race. The two unstructured tasks were to write a position paper on the church's stand on (*c*) legalizing abortion and (*d*) a "Black Caucus" within the Unitarian-Universalist church. Leader position power was low.

Mitchell (1969) originally computed correlations for groups with poor as well as with good leader-member relations, and these correlations were included in the Graen et al. (1970) critique. The Mitchell study was not, however, designed as a test of the model, and an analysis of the data showed that only 2 of the 64 group sessions in this study had group atmosphere scores below 55, a score which is roughly at the median of group atmosphere scores for similar studies. As pointed out before, a finding of this type is neither unusual nor unexpected in ad hoc groups. Most researchers, testing the model, have, therefore, tried to increase the difference between groups with high and low group atmosphere scores by using only the upper and lower thirds of the distribution, whenever this was possible. In light of the high group atmosphere scores in the Mitchell study, all groups were classified as falling into Octants II and IV depending upon task structure. The resulting rank-order correlations for Octant II were .24 and .17 ($n = 16$), and for Octant IV, .43 and .38 ($n = 16$).

Executive development workshop. An almost identical study was conducted by Fiedler as part of an executive development program. Here again, all but one of the group atmosphere scores was above the usual cutting score, and all groups were classified as having high group atmosphere. The groups were given a relatively structured task and an unstructured task (i.e., routing a truck convoy and writing a recruiting statement inviting college students to become junior executives). All leaders were rated as having low position power, and thus all groups fell into Octants II and IV. The correlations were .34 and .51 for the structured and unstructured tasks, respectively (ns = 11). (See also the section reviewing results on groups in training which might apply to the church leadership and the executive development workshop studies.)

West Point Cadets. Skrzypek[6] recently completed a study of 32 four-man groups composed of cadets at West Point. Leaders were chosen from among a pool of 400 men whose LPC scores fell either one standard deviation above or below the mean. Members were assigned at random. Unlike other laboratory studies, which used postsession questionnaires, leader-member relations in this study were determined a priori on the basis of previously obtained sociometric ratings that identified well-accepted and not-accepted leaders among the cadets. Position power was varied by informing the group that the leader in the high position power would be evaluating each of his group members at the conclusion of the tasks and that these evaluations would become part of the individual's leadership score (a very important aspect of West Point's system). The leaders in the low position power were instructed to act as chairmen, and the group was told that the members as a group would be evaluated on their performance.

---

[6] Skrzypek, G. J. The relationship of leadership style to task structure, position power, and leader-member relations. West Point, New York: United States Military Academy, 1969.

Each group performed one structured task and one unstructured task in counterbalanced order. The structured task consisted of drawing a plan for barracks and a military-post area to scale. The unstructured task required the groups to design a program which would educate enlisted men in overseas assignments on world politics and maintain their interest throughout their tour.

The results were as follows for each of the eight octants: Octant I, − .43; Octant II, − .32; Octant III, .10; Octant IV, .35; Octant V, .28; Octant VI, 13; Octant VII, .08; Octant VIII, − .33. Thus, while none of the correlations reached the .05 level of significance, all but the correlation for Octant III were in the predicted direction.

Nonclassifiable Studies.

Student nurses. A study by Reilly[7] was excluded because the judges could not agree with one another or with the investigator on the position power of the leaders. Reilly studied groups composed of nurses who were given successively structured and unstructured discussion problems. Whether the leader actually had high position power was questioned because the leader was a fellow student, and her additional responsibilities were limited: They consisted of making certain arrangements for the groups as well as assigning 20% of the grade each student would obtain. This study is discussed in a later section, since it may also be classified as a training study.

Experimental change of position power. A second nonclassifiable study was conducted by Nealey and Shiflett (S. Nealey, personal communication, 1969) who attempted experimentally to induce a change of situational favorableness from Octant III to Octant IV and vice versa. The groups were assembled on the basis of LPC and intelligence scores. The experimental manipulation, changing the groups, did not succeed although the study prior to the experimental manipulation yielded relations between LPC and group performance. Here again, the judges did not agree on whether the leader's position was, in fact, strong, since in the appropriate experimental condition the leader's position power was established basically by instructing the leader to role play a person with a powerful or weak position power.

Graen, Orris, and Alvares (1971) described two laboratory experiments which were specifically designed to test the contingency model. These studies also constituted the main basis for a recent article in this journal by Graen et al. (1970), which questioned the adequacy of the model.

Graen et al. (1971) used 78 and 96 male college students randomly assembled into 52 and 64 three-man groups, respectively. Each group worked on two tasks. The students were paid for their participation, and one person in each three-man group was chosen at random to serve as the leader during the first task. Another member of the same group served as leader during the second task.

Each of the groups was given one structured and one unstructured task in counterbalanced order. Using an 8-point scale for assessing task structure (Fiedler, 1967a, p. 25ff) Graen et al. (1971) chose two structured and two unstructured tasks. Position power was varied by giving the leaders of one set of groups

> superior formal status relative to the members . . . special information about the task, and . . . the highest decision-making authority and responsibility. In the weak leader condition, the leader role was one of discussion leader without special information and with decision-making authority and responsibility close to that of the members. Each group worked on both tasks under only one of the two power conditions. The task sequence was randomized within power conditions [Graen et al., 1971, p. 198].

Group atmosphere scores were obtained after each task session, and the groups were divided into those whose group atmosphere scores had fallen above and below the median for the entire set. Groups were then assigned to the appropriate cell of the contingency model by dichotomizing the three situational variables of position power, task structure, and group atmosphere. The results of the two experiments are given in Table 4.

Graen and his associates claim that their two studies followed as closely as possible the prescriptions of the contingency model and

---

[7] Reilly, A. J. The effects of different leadership styles on group performance: A field experiment. Ames, Iowa: Industrial Relations Center, Iowa State University, 1968.

TABLE 4

Correlations between Leader LPC and Performance in the Graen et al. Studies

Experiment	Octant							
	I	II	III	IV	V	VI	VII	VIII
1	.47[a]	−.41	.46[a]	.33	.25	−.39	.43	−.33
2	−.13	.18[a]	.02[a]	−.08[a]	−.52[a]	−.43	.45	.44[a]

[a] Correlations in a direction counter to the hypothesis of the model.

research methodology used in its development. Because their results were nonsignificant, the authors concluded that their studies, therefore, "cast doubt on the plausibility of the contingency model [Graen et al., 1971, p. 201]."

Nonsignificant results obviously can occur for any number of reasons, including inadequate experimental design. A study that attacks a theory, therefore, must not only be methodologically sound (especially if it does not support an alternative hypothesis), but it must also guard against the possibility of obtaining nonsignificant and randomly distributed data because of inappropriate or marginal experimental manipulations.

A methodological critique in the *Journal of Applied Psychology* (Fiedler, 1971) pointed to various weaknesses in the manipulation of three independent variables in the Graen et al. (1971) experimental design. The manipulations were inadequate and make the studies inappropriate for inclusion in the present analysis. A summary of the inadequacies is presented below. The discussion is based on data in the original Graen et al. (1971) paper as well as subsequent data included in a rejoinder to the methodological critique of their studies.

1. *Task structure.* Task structure was assessed by means of a rating scale with scores ranging from a maximum of 8.0 to a minimum of 1.0 (Fiedler, 1967a). The average task structure scores of the studies on which the original contingency model paper was based averaged 7.39 for the structured and 3.15 for the unstructured tasks. In contrast, the scores in the Graen et al. (1971) studies were 5.85 and 5.45 for the two structured tasks, and 3.69 and 3.60 for the two unstructured tasks. The scores for the structured task were, therefore, less than 1 scale point above the cutting score of 5.0, and the differences between structured and unstructured tasks were only 2.17 and 1.85 scale points. This clearly represents a very weak manipulation as compared with the differences of 4.24 in the original tasks.

While it may be argued that the task structure scores in some of the original studies were similar to those in the Graen et al. studies, an experiment which seeks to disconfirm a theory should not use marginal manipulations in testing the null hypothesis.

2. *Position power.* Graen et al. (1971) manipulated this variable by giving the randomly chosen leader "superior formal status" and "special information about the task," and delegating to him "the highest decision-making authority and responsibility." This was done here by talking directly to the leader in the presence of the members, presenting written task instructions to the leader only, "maintaining body orientation and eye contact with the leader," and giving an official timing device to the leader. In the weak position power condition, the formal status of the leader was not reinforced, verbal instructions were addressed to the group, written instructions were given to no one in particular, and the timing device was placed in the center of the table.

A number of reasons suggest that this power manipulation is inadequate. First of all, it is difficult to believe that the particular experimenter behaviors—like looking the leader in the eye, and giving the leader of an ad hoc group a timing device and written instructions—are sufficiently powerful manipulations to do all the things the authors expected to accomplish. Position power is conceptualized as providing the leader with some real power to give rewards and sanctions. In other words, the leader must have some fate control over his members. It is very difficult to give high position power to a leader in any laboratory situation. Where this was done successfully, it was usually accomplished by using individuals who had some formal position power outside the laboratory. Thus, the Belgian Navy study (Fiedler, 1966) compared petty officers with recruit leaders; a study of ROTC cadets by Meuwese and Fiedler (Fiedler, 1967a) and a study of West Point cadets by Skrzypek (see Footnote 6) used cadet officers who held higher rank than their members. Moreover, Graen et al. (1971) demoted the leader appointed for the first task session to member status in the second session and made another member of the same group the leader in the second session. This procedure is likely to dilute the formal leadership power.

Graen et al. (1971) subsequently reported data on perceived leader influence in support of their claim that the position power manipulation actually had been effective. Cell means of perceived influence ratings are reported in Table 5.

Thus, the difference in means over octants in Experiment 2 was only .16. A *t* test comparing the perceived influence ratings in the strong and weak position power conditions was 1.746, which is not significant. As can be seen, some of the weak position power octants had means which were higher than some of the supposedly high position power octants. Specifically, two of the "weak" octants had mean scores of 5.38, while two of the "strong" octants in the first experiment had mean scores of only 5.14. The position power manipulation in Experiment 2 appears, therefore, not to have been effective.

3. *Distribution of LPC scores.* Leaders were assigned at random and LPC scores were obtained after the three-man groups had been assembled. This procedure does not properly assure that leader LPC scores within each octant will have similar means and distributions. It is obvious that a meaningful test of the contingency model cannot be obtained if, for example, all groups within one octant have leaders with high LPC scores, while all groups in another octant have leaders with very low scores. An appropriate test is an analysis of variance to determine whether the means of various octants are reasonably similar; if so, the *F* test will be nonsignificant.

TABLE 5

CELL MEANS OF PERCEIVED INFLUENCE RATINGS IN GRAEN ET AL. STUDY

Condition	Octant				$\bar{X}$
Experiment 1					
Strong position power	5.83	6.17	5.14	5.14	5.57
Weak position power	4.33	4.83	3.24	4.86	4.32
M_{diff}					1.25
Experiment 2					
Strong position power	5.75	5.63	6.13	3.50	5.29
Weak position power	5.38	5.00	5.38	4.75	5.13
M_{diff}					.16

Data provided by Graen et al. (1971) show that the means of LPC scores within the various octants ranged from 84.7 to 42.7 in the first experiment and from 77.3 to 52.9 in the second experiment. A one-way analysis of variance was performed for each study with the eight octants as cells in the design. The *F* ratio for the second study was not significant (1.17), indicating that the various octants did not differ in mean LPC. The *F* ratio for the first study was, however, 3.10, which is significant at the .01 level and indicates that the distribution of leader LPC scores differed markedly from octant to octant. In other words, some octants, (e.g., Octant IV) contained few, if any, low LPC leaders, while others (e.g., Octant V) contained few, if any, high LPC leaders.

Since the position power manipulation in addition to the weak task structure manipulations in Experiment 2 and the LPC score distributions in Experiment 1 were inadequate for testing the contingency model, neither experiment was included in the validation analysis.

Summary of Results

Table 6, which summarizes all correlations from acceptable studies, shows that the median correlations for six of the seven octants are in the predicted direction (Octant VI was not predicted). Of these, the joint probabilities of the correlations in Octants I, III, and IV are significant below the .05 level. Also, 34 of the 45 correlations are in the predicted direction, a finding significant at the .01 level by binomial

TABLE 6

SUMMARY OF FIELD AND LABORATORY STUDIES TESTING THE CONTINGENCY MODEL

Study	Octants							
	I	II	III	IV	V	VI	VII	VIII
Field studies								
Hunt (1967)	−.64		−.80		.21		.30	
	−.51		.60				−.30	
Hill (1969)[a]		−.10	−.29			−.24	.62	
Fiedler et al. (1969)		−.21		.00		.67*		−.51
O'Brien et al. (1969)		−.46		.47		−.45		.14
Laboratory experiments								
Belgian Navy	−.72	.37	−.16	.08	.16	.07	.26	−.37
	−.77	.50	−.54	.13	.03	.14	−.27	.60
Shima (1968)[a]		−.26		.71*				
Mitchell (1969)		.24		.43				
		.17		.38				
Fiedler exec.		.34		.51				
Skrzypek[a]	−.43	−.32	.10	.35	.28	.13	.08	−.33
Median								
All studies	−.64	.17	−.22	.38	.22	.10	.26	−.35
Field studies	−.57	−.21	−.29	.23	.21	−.24	.30	−.33
Laboratory experiments	−.72	.24	−.16	.38	.16	.13	.08	−.33
Median correlations of Fiedler's original studies (1964)	−.52	−.58	−.33	.47	.42		.05	−.43

Note.—Number of correlations in the expected direction (exclusive of Octant VI, for which no prediction had been made) = 34; number of correlations opposite to expected direction = 11; p by binomial test = .01.

[a] Studies not conducted by the writer or his associates.

* $p < .05$.

test. It should be noted that the number of correlations in the predicted direction would be significant at the .01 level by binomial test even if we included the two Graen et al. (1970, 1971) experiments. These results permit the conclusion that we are not dealing with random effects: Group performance appears to be contingent upon leadership style and situational favorableness.

On the other hand, 5 of the 10 correlations obtained in Octant II are in the positive rather than in the negative direction. These counterexpectational positive correlations, although found only in this octant, throw considerable doubt on the overall generality of the relationship predicted in the Fiedler 1964 paper. It is, therefore, essential that we examine the relations in greater detail.

It will be recalled that in the 1964 study, data for Octants I, II, and V came from field studies, while the data for Octants III, IV, VII, and VIII came, with but one exception, from laboratory studies. Field and laboratory results should, therefore, be examined separately.

Field studies. The median correlations for field studies are quite similar to those predicted in the Fiedler 1964 paper. All of the medians are in the predicted direction, and 13 of the 15 predicted correlations are in the expected direction, which is significant at the .05 level. The curve, based on relatively few studies, is not as regular as that obtained in 1964. Octant VI was not predicted. The predicted median correlation of .05 in Octant VII which we based on 12 correlations in the 1964 study, now is shown with a median correlation of .30, based on three correlation coefficients. Overall, considering the small number of studies and the small number of cases within each of these studies, the results seem rather remarkably consistent with the 1964 data, suggesting that the model is valid for the prediction of leadership performance under field conditions.

Experiments and laboratory studies. Only the Belgian Navy, the West Point, and Graen et al. (1971) studies provided data for all predicted octants. It is, therefore, hazardous to draw more than tentative conclusions for any octants but II and IV for which sufficient data are available. Data for Octants I and IV tended to support the prediction of the contingency model, while they did so only directionally in Octants III, V, VII, and VIII. The data clearly indicate that the model does not adequately predict leadership performance in Octant II of laboratory studies. On the other hand, 22 of the 29 predicted correlation coefficients were in the expected direction, which is significant at the .01 level for binomial tests.

A number of possible explanations for the nonpredicted results suggest themselves. The most parsimonious and obvious of these might simply be that it is difficult to manipulate leadership variables in laboratory experiments, and some important aspects of real-life situations do not permit themselves to be readily built into the laboratory.

EXTENSIONS OF THE CONTINGENCY MODEL

A number of studies have tested the more general hypothesis that the situational favorableness affects the relationship between leadership style and performance. Some of these studies tested groups in situations ranging from the very favorable to the very unfavorable; other studies considered groups falling on only two points on the situational favorableness continuum. The various group situations were categorized as being favorable, intermediate, or unfavorable situations by two or more judges. However, since the degree of situational favorableness in some of these investigations was not operationally specified in advance, the rejection of the null hypotheses becomes correspondingly more hazardous in these cases. This is especially so in studies in which the statistical relations do not reach the commonly accepted level of significance. On the other hand, it is difficult to obtain a large sample of groups or organizations in any one study; and it is essential, therefore, that we consider the cumulative evidence from different investigations.

Experimental variation in leadership behavior. Shaw and Blum (1966) instructed nine leaders to act in a highly controlling and directive manner, while a second set of 9 leaders was told to be permissive and passive. The groups were given three tasks in counterbalanced order, and the tasks varied in degree of structure. Permissive, passive leaders performed more effectively in the two relatively unstructured tasks (roughly Octant IV), the directive leaders, as predicted by the investigators, performed more effectively in the highly structured tasks (roughly Octant II). These results conform to the general expectations of the model.

Objectification of situational favorableness by structural role theory. O'Brien (1969) measured situational favorableness objectively by applying methods of structural role theory (Oeser & Harary, 1962, 1964). O'Brien assumed that a situation would be more favorable to the leader the greater the number of paths he had to the task structure. The more readily and directly the leader could influence task performance, the greater his influence and control of task-relevant group behavior.

The structural role theory deals with relations among three elements, namely, persons, positions, and tasks. Relations among the first indicate the interpersonal relationships, relations among the positions indicate authority relations, and relations among the tasks indicate allocations and task sequences. O'Brien assembled groups on the basis of personal compatibility depending upon the members' similarity or dissimilarity in scores on Schutz' FIRO scale (1958). The group task consisted of constructing models from spheres and sticks according to a given pattern. Position and task allocation were manipulated by determining the means by which the leader could interact in the task performance. A coefficient of situational favorableness could then be computed which, in oversimplified form, expressed the leader's paths to the task structure as a ratio of all possible paths. The higher the ratio, the more favorable the leader's position power.

The results of O'Brien's study, based on the correlation between leader LPC and the number of models produced, showed the predicted relationships. For 16 groups with high situational favorableness the correlation was − .08; for 16 groups with intermediate favorableness it was .77 (significant at .01); and for 32 groups with low situational favorableness the correlation was − .13.

Heterocultural American and Indian groups. Anderson (1966) conducted a laboratory experiment which used graduate students from India as well as from the United States. These groups consisted of one American leader, one American group member, and one Indian group member. The tasks required the groups to negotiate an agreement on hiring practices between an Indian village and an American company, and to compose two different stories based on the same TAT card. Half the leaders had high LPC, and half had low LPC scores. Of these, half were instructed to be as considerate as possible in their leadership behavior, while the other half were told to structure the situation as firmly as possible.

Anderson performed a post hoc analysis that scaled the situations on the basis of favorableness. This scaling suggested that the TAT tasks were more favorable than the hiring problem (which required negotiation) and that the considerate condition would be more favorable than the structuring condition. This ordering of situations was supported by leader responses to questions about their anxiety and tension in each of these conditions.

The correlations for the four tasks in order of favorableness (TAT-considerate, TAT-structuring, negotiation-considerate, negotiation-structuring) were − .50, .21, − .22, and − .12, thus suggesting a curvilinear relationship between LPC-performance correlations and situational favorableness (*n*s = 8).

Chemical processing companies. Lawrence and Lorsch (1967) compared the performance of six chemical processing companies, each of which had four subsystems: production, sales, applied research, and basic research. Each subsystem was under the direction of a senior management official, either at the vice-presidential level or immediately below. The structure of these subsystem tasks was rated by a number of judges. Fundamental research was rated as lowest in structure, while production was rated as highest. This reviewer correlated the manager's LPC and rated performance from the Lawrence and Lorsch data. The results follow the expectation that a low LPC manager would perform better on structured tasks, while a high LPC manager would perform better in unstructured situations. The correlations, in order of rated structure, were for production, − .50; sales, − .31; applied research, − .10; and fundamental research, .66. (*n*s = 6).

Since these results were obtained without reference to leader-member relations, they suggest that the leader-member relations dimension might be relatively less important at higher levels of the organization. This seems reasonable since the manager at the third and higher levels usually has very few direct contacts with production workers at the nonsupervisory levels, and relatively few contacts with first-line supervisors.

Psychiatric nursing organization. Nealey and Blood (1968) investigated the psychiatric nursing service of a large Veterans Administration hospital. LPC scores were obtained from supervisors at the first and second levels of the organization, namely, head nurses in charge of a ward, and "unit supervisors" in charge of one of the hospital's six large units, which are both quite comparable as to structure and function. Performance of wards and of units was judged by management personnel at the level above the head nurse and the unit supervisor, respectively.

The most important difference between the job of the head nurse and that of the unit supervisor for purposes of the present analysis appears to be the structure of her subordinates' tasks. The work of the psychiatric aide is relatively structured since there are fairly specific guidelines available on the management of psychiatric patients and ward personnel. The job of the unit supervisor, or more precisely, the work of the head nurses she supervises, requires considerably more policy and decision making, and it is correspondingly less structured. Nealey and Blood correlated the LPC scores of these nursing supervisors with rated performance. In a very similar study, Nealey and Owen (1970) about 18 months later recomputed these correlations on 25 head nurses,

TABLE 7

Summary of LPC–Group Performance Correlations of Studies Extending the Contingency Hypothesis

Study and date	Favorable			Intermediate			Unfavorable		
O'Brien (1969)	−.08			.77[a]				−.13	
Anderson (1966)	−.50		.21		−.22			.12	
Lawrence & Lorsch (1967)[a]	−.50	−.10	−.13	.66					
Nealey & Blood (1968)	−.22			.79[a]					
Nealey & Owen (1970)	−.50[a]								
Fiedler & Barron[b]									
Task I	−.42	−.56	−.32	.67	−.08	−.01	−.53	−.72	.18
Task II	−.71	−.59	.69	.41	−.15	−.20	−.47	−.61	−.14
Median		−.37			.20			−.30	

Note.—The location of the correlation coefficient in the table indicates degree of judged favorableness of the leadership situation. The farther to the left, the more favorable the situation.

[a] Study not conducted by writer or his associates.

[b] In Fiedler (1967a).

15 of whom had participated in the first study. The correlations between head-nurse LPC and performance for the first and second studies were − .22 ($n = 21$) and − .50 ($n = 25$, $p < .05$), and for unit supervisors tested in the first study, .79 ($n = 8$, $p < .01$). Thus, as in the Lawrence and Lorsch study, the structure of the supervisory task strongly moderates the direction of the relationship between LPC and organizational performance.

Stress as an index of situational favorableness. A study by Fiedler and Barron (cited in Fiedler, 1967a) investigated the relationship between leader LPC and creative group performance under varying conditions of stress. Fifty-four three-man teams composed of ROTC cadets participated in this study under relatively stress-free conditions, intragroup conflict, and relatively severe external stress. Within each of these stress conditions the groups were divided into those with high, medium, and low leader group-atmosphere scores. The resulting LPC-performance relations were then plotted against stressfulness of the situation and resulted in a bow-shaped curve with low LPC leaders performing better than high LPC leaders in relatively stress-free and stressful conditions, and high LPC leaders performing better in situations of intermediate stress.

Summary of studies testing the contingency theory. The results presented in this section are, in one respect, quite weak, and in another, quite strong. The investigations do not permit an exact test of a theory since the methodology, the criteria, and the subject populations vary from study to study. At the same time, these studies provide consistent cumulative evidence that the correlation between leadership style and group performance is moderated by the situational favorableness dimension even though this dimension is operationalized in a wide variety of ways. Thus, the Shaw and Blum experiment, the Lawrence and Lorsch investigation, as well as the Nealey and Blood and the Nealey and Owen studies show that the structure of the task is important in determining the direction of the correlation between LPC and group performance. The Fiedler and Barron (Fiedler, 1967a) study used stress as an index of situational favorableness. The O'Brien experiment presented an ingenious new metric of situational favorableness based on structural role theory, and the results of his study yielded relations in the expected direction with one of the correlations highly significant. Finally, the Anderson (1966) experiment, based on post hoc analysis of the data, also yielded data that tended to show an interaction of task and leadership style. Thus, notwithstanding the diversity of the studies, there is clear evidence that identified situational components determine, in part, the type of leadership style that a particular group requires for effective performance.

EXTENSION OF THE MODEL TO COACTING GROUPS

A number of field studies have attempted to investigate the relation of leader LPC scores to the performance of coacting groups and organizations (Fiedler, 1967b). The number of studies reported is now sufficient so that a review of the findings appears appropriate.

Task Groups and Organizations

Craft shops and grocery markets. Two organizations studied by Hunt (1967) were coacting. These were craft shops in the physical science laboratory and grocery departments of chain stores. Both of these types of organizations performed coacting tasks that were highly structured. In both organizations the leader's position power was high. Performance of craft shops was assessed by ratings, while grocery departments were evaluated on the basis of an objective measure of number of man-hours over total sales volume.

The correlations for six workshops with high, and five with low group atmosphere scores (comparable to Octants I and V), were, respectively, $-.48$ and $.90$ ($p < .05$), those for 13 supermarkets with high group atmosphere, $-.06$, and those with low group atmosphere, .49.

Hospital departments. Hill's (1969) study included various departments in a hospital, including the nursing service, the controller's department, dietetics, housekeeping, stores, central supply, and maintenance. All were rated to be coacting, and all supervisory personnel were rated as having high position power. The nursing service was rated by Hill as having an unstructured task, while other departments were rated as having structured tasks. Classifying the groups on the basis of the contingency model classification would make the nursing service fall into Octants III and VII and other departments into Octants I and V, depending on the leader's group atmosphere score. The correlations for the Octant I and V groups were $-.21$ ($n = 8$) and $.52$ ($n = 8$) and for the Octant III and VII groups (nursing) $-.32$ ($n = 7$) and $.87$ ($n = 7$, $p < .05$).

Telephone offices. Bates (1967) conducted a study of telephone supervisors in two offices of the Bell Telephone Company. The total sample consisted of 112 operators working in 13 groups. Of these employees, 21 were first-line supervisors who provided LPC and group atmosphere scores. The task was rated as highly structured, and the position power as high. These groups would then fall into Octants I and V.

Bates obtained two criterion scores, a "quality index" reflecting the accuracy and courtesy with which the operators handle calls as well as the accuracy in billing calls. A second index, called the "load factor," reflects the average number of calls handled by an office. Three telephone executives who were independently interviewed agreed that the quality index is the more important and that offices rarely performed below the minimum load factor. Since the quality and load indexes were negatively correlated ($-.43$), only the quality index was here considered. Bates divided his groups into six with relatively high group atmosphere and seven with low group atmosphere scores. The corresponding correlations between supervisors' LPC scores and quality indexes were $-.77$ and .75.

School principals. McNamara (1967) investigated the effectiveness of elementary school principals in 32 elementary schools of the Edmonton, Alberta, school system. The schools were relatively small in size, some containing as few as six teachers. Although McNamara spoke of his schools as interacting, all four judges classified the schools as coacting. Their reason was that there is relatively little interaction among teachers in the performance of their instructional duties, nor is it likely that the principal's job demands that he share his administrative duties with his staff or teachers. The principal's position power was rated as being high.

The effectiveness of schools was judged by five members of the school system's administrative staff. Leader-member relations were indexed by group atmosphere scores, with the upper third of the schools and the lower third of the schools used, and the middle third deleted from the sample.

The correlation between principal's LPC and performance in high group atmosphere schools (Octant I) was $-.48$ ($n = 11$), while that in low group atmosphere schools (Octant V) was .31 ($n = 12$).

Training Situations

Naval aviation cadets. Fiedler and Hutchins (Fiedler, 1967a) conducted a study of aviation cadets (including some commissioned officers in flight training) and their instructors. The student pilots were in the advanced course which concentrated on formation flying, a rather anxiety-arousing phase of training. The cadets were assigned to "flights" of eight men, and each flight was under the direction of an instructor team. The performance of the men in each flight was evaluated by instructors regularly assigned to another squadron of the training base.

Two samples of 16 instructors were tested, using their groups' performance scores as criterion of leadership (Octants I and V). The correlation of head instructors' LPC scores and performance of their flights was .45 and .17. Fiedler (1967a) also obtained correlations between the sociometrically chosen men, that is, informal leaders with very little position power. Since all were sociometrically chosen, all had good leader-member relations. The correlation between the LPC scores of 22 leaders and performance ratings was .55 ($p < .02$) and .28 for *ns* of 22 and 15. The lower correlations in the second set of teams may be due to the smaller range and standard deviation of performance scores in the second sample of teams.

Management trainees. A study by Seifert (1969) used 14 management training groups in Germany. These groups, each consisting of 25 to 30 participants, remained together for almost 1 year. Similar to the naval aviation cadet study, each group had a team of instructors and a head instructor. Each group also had an informal leader, designated as spokesman of the group, but having very little position power. Similar to the naval aviation cadets, Seifert reported that the men were under considerable pressure and anxiety lest they fail. (Although Seifert treated these groups as interacting, all four judges considered the Seifert study one of coacting groups.) Effectiveness was rated by instructors and was based on several scales of motivational and performance characteristics of the men.

Correlations between the LPC of head instructors and performance was .56 for those with high group atmosphere and −.20 for those with low group atmosphere scores. The correlation between spokesmen's LPC and performance scores was .45, regardless of group atmosphere.

Student nurses. A study by Reilly, mentioned above (see Footnote 7), was designed to test the contingency model. However, the judges rating Reilly's method were divided on the leader's position power. The study involved the assembly of 14 groups of student nurses with each group required to complete 10 sets of discussion problems and examination questions posed by the faculty of the school over the course of the year. Some of these problems were highly structured; that is, there was an answer or a solution available. Other problems were highly unstructured; that is, the group had to discuss a case or an ethical issue for which no definite answer was available. Reilly reported a correlation of .63 between leader LPC and group performance, irrespective of leader group atmosphere scores.

The leader had responsibility for arranging the sessions and seeing that the group's solutions were typed and handed in to the faculty. She also had the responsibility to assign 20% of each of her fellow students' grades. Leaving aside for the moment the question of position power on which our judges disagreed, or that these groups may well have been interacting in most tasks, the more important point might well be that the Reilly study, as well as the naval aviation cadet study by Fiedler and Hutchins (Fiedler, 1967a) and the Seifert study of German management trainees, is distinguished by having as its purpose the training of the group members rather than the performance of some task that results in an output beneficial to the organization.[8] In all of the studies, the exercise was designed to be beneficial to the individual group member. There might well be a psychological difference in leading a group for the purpose of benefiting the members and leading a group for the pur-

---

[8] An earlier unpublished study of classroom performance by Marse (1958) showed negative correlations of $-.70$ ($n = 6$) and $-.36$ ($n = 12$) for physics and rhetoric section instructors at the University of Illinois where rated student performance was the criterion. In that study, however, the class members never interacted in groups as part of their training. Whether this is an important determining factor remains to be seen.

TABLE 8

CORRELATIONAL SUMMARY OF STUDIES EXTENDING THE CONTINGENCY MODEL TO COACTING TASK AND TRAINING GROUPS

Group	Octants							
	I	II	III	IV	V	VI	VII	VIII
Task groups								
Craft shops	−.48				.90[a]			
Groceries	−.06				.49			
Hospital departments[a]	−.21		−.32		.52		.87	
Telephone offices	−.77				.75			
School principals[a]	−.48				.31			
Median	−.48		−.32		.52		.87	
Training groups								
Naval aviation								
Chief instructors	.45 .17							
Informal leaders			.55 .28					
Management trainees[a]								
Head instructors	.56							
Informal leaders	.45					−.20		
Student nurses[a]			.45 .63					
Medians	.45		.50			−.20		

[a] Studies not conducted by writer or his associates.

pose of benefiting the organization. Our data suggest that this may be the case.[9]

Summary of results from coacting groups. As we suggested in the discussion of the Reilly (see Footnote 7) study, the results obtained from studies of coacting groups can best be discussed by separating results from task groups and results from training groups (Table 8).

The results from task groups are highly consistent in showing that groups or organizations with structured tasks and high leader position power perform more effectively under low LPC leaders when the group climate is rated as favorable, and more effectively under high LPC leaders when the group climate is unfavorable. Classifying coacting groups in the same manner as interacting groups indicates that the correlations are quite similar in the two octants for which data are available, namely, Octants I and V.

Hill, in his study of a hospital, classified the supervising nurses' job as unstructured. The correlation for Octant III is in the predicted direction, and the correlation for Octant VII, while much higher than would be expected for this octant, is not incompatible with data from interacting groups.

The data on coacting *task* groups suggest that the distinction between interacting and coacting task groups might be unnecessary, while the distinction between *task* groups and *training* groups might be essential. The latter appear to follow quite dissimilar rules.

The Seifert (1969), the Fiedler and Hutchins (Fiedler, 1967a), and Reilly (see Footnote 7) studies suggest that groups in training might constitute a valid subclassification for leadership studies, a classification that needs to be intensively examined further. The possibility should be considered that this set of studies might also include those by Mitchell (1969) and Fiedler which used participants in leadership workshops.

Recent empirical findings from research on leadership training. Despite intensive efforts, research has failed to show that leadership experience or leadership training systematically improve organizational performance (Campbell et al., 1970). These disappointing results can be deduced from the contingency model. Specifically, we can interpret leadership experience and training as improving situational favorableness (e.g., human relations training is supposed to improve leader-member relations; technical training would make the task appear more structured). Training and experience should then differentially affect the performance of high and low LPC leaders. The contingency model would then predict that training for intermediate situations will improve the performance of high LPC leaders—but *decrease* that of low LPC leaders. Likewise, training and experience for favorable and unfavorable situations will improve performance of low LPC leaders—but *decrease* performance of high LPC leaders. These results have now been obtained in several studies and further support the contingency model (Fiedler, in press).

---

[9] Note, however, that a school principal's group membership consists primarily of teachers and clerical staff, not of students.

SUMMARY

A review of studies testing and extending the contingency model evaluated the model's predictive power in field and laboratory situations and in coacting task and training groups.

The model seems to predict leadership performance in field situations, but not completely in laboratory situations. The major discrepancy between the model and the field studies on the one hand, and the laboratory studies on the other, was in Octant II, where the model predicted negative correlations, while the laboratory studies showed predominantly positive correlations between LPC and group performance. This discrepancy, if it is not due to chance, may well bring to light important aspects of leadership interactions that are not usually reproduced in laboratory situations.

A series of studies, extending the theory, was reviewed. Taken as a group, these studies provide strong evidence that the situational favorableness dimension does indeed moderate the relationship between leadership style and group performance, and that it therefore provides an important clue to our understanding of leadership phenomena.

Finally, studies of coacting groups and organizations suggested that we differentiate between groups that exist primarily for the benefit of the organization—that is, the typical task groups—and groups that exist for the benefit of the individual—groups of trainees. Coacting task groups appear to follow the predictions of the contingency model, at least for Octants I and V. Data from groups-in-training showed consistently positive relations between leader LPC and group performance measures in Octants I and III, and only one small negative correlation in Octant VI. These findings suggest a new approach to the classification of coacting groups and training groups which might lead to a better understanding of managerial performance in task and training organizations.

REFERENCES

Anderson, L. R. Leader behavior, member attitudes, and task performance of intercultural discussion groups. *Journal of Social Psychology*, 1966, **69**, 305–319.

Bates, P. A. Leadership performance at two managerial levels in the telephone company. Unpublished bachelor's thesis, University of Illinois, 1967.

Butterfield, D. A. An integrative approach to the study of leadership effectiveness in organizations. Unpublished doctoral dissertation, University of Michigan, 1968.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., & Weick, K. E. *Managerial behavior, performance, and effectiveness.* New York: McGraw-Hill, 1970.

Fiedler, F. E. A contingency model of leadership effectiveness. In L. Berkowitz (Ed.), *Advances in experimental social psychology*. New York: Academic Press, 1964.

Fiedler, F. E. The effect of leadership and cultural heterogeneity on group performance: A test of the Contingency Model. *Journal of Experimental Social Psychology*, 1966, **2**, 237–264.

Fiedler, F. E. *A Theory of leadership effectiveness*. New York: McGraw-Hill, 1967. (a)

Fiedler, F. E. Führungsstil und Leistung koagierender Gruppen. *Zeitschrift für experimentelle und angewandte Psychologie*, 1967, **14**, 200–217. (b)

Fiedler, F. E. Note on the methodology of the Graen, Orris, and Alvares studies testing the contingency model. *Journal of Applied Psychology*, 1971, **55**, 202–204.

Fiedler, F. E. *Leadership*. (A module for the General Learning Press, in press.

Fiedler, F. E., O'Brien, G. E., & Ilgen, D. R. The effect of leadership style upon the performance and adjustment of volunteer teams operating in a stressful foreign environment. *Human Relations*, 1969, **22**, 503–514.

Fishbein, M., Landy, E., & Hatch, G. Consideration of two assumptions underlying Fiedler's Contingency Model for the prediction of leadership effectiveness. *American Journal of Psychology*, 1969, **4**, 457–473.

Graen, G., Alvares, K., Orris, J. B., & Martella, J. A. Contingency model of leadership effectiveness: Antecedent and evidential results. *Psychological Bulletin*, 1970, **74**, 285–296.

Graen, G., Orris, J. B., & Alvares, V. M. Contingency model of leadership effectiveness: Some experimental results. *Journal of Applied Psychology*, 1971, **55**, 196–201.

Hill, W. The validation and extension of Fiedler's theory of leadership effectiveness. *Academy of Management Journal*, 1969, March, 33–47.

Hunt, J. G. Fiedler's leadership contingency model: An empirical test in three organizations. *Organizational Behavior and Human Performance*, 1967, **2**, 290–308.

Kretzschmar, V., & Luecke, H. E. Zum Fiedlerschen Kontingenzmodell Effektiver Führung. *Arbeit und Leistung*, 1969, **23**, 53–55.

Lawrence, P., & Lorsch, J. Differentiation and integration in complex organizations. *Administrative Science Quarterly*, June 1967, **12**, 1–47.

Marse, J. E. Assumed similarity between opposites and the performance of leadership functions. Unpublished master's thesis, University of Illinois, 1958.

McNamara, V. D. A descriptive-analytic study of directive-permissive variation in the leader behavior of elementary school principals. Unpublished master's thesis, University of Alberta, 1967.

Mitchell, T. R. Leader complexity, leadership style, and group performance. Unpublished doctoral dissertation, University of Illinois, 1969.

Mitchell, T. R. Leader complexity and leadership style. *Journal of Personality and Social Psychology*, 1970, **16**, 166–174.

Nealey, S. M., & Blood, M. Leadership performance of nursing supervisors at two organizational levels. *Journal of Applied Psychology*, 1968, **52**, 414–422.

Nealey, S. M., & Owen, T. M. A multitrait-multimethod analysis of predictors and criteria of nursing performance. *Organizational Behavior and Human Performance*, 1970, **5**, 348–365.

O'Brien, G. E. Group structure and the measurement of potential leader influence. *Australian Journal of Psychology*, 1969, **21**, 277–289.

Oeser, O. A., & Harary, F. A mathematical model for structural role theory I. *Human Relations*, 1962, **15**, 89–109.

Oeser, O. A., & Harary, F. A mathematical model for structural role theory II. *Human Relations*, 1964, **17**, 3–17.

Schutz, W. C. *FIRO: A Three-dimensional theory of interpersonal behavior*. New York: Holt, Rinehart & Winston, 1958.

Seifert, K. H. Untersuchungen zur Frage der Führungseffektivität. *Psychologie und Praxis*, 1969, **13**, 49–64.

Shaw, M. E., & Blum, J. M. Effects of leadership style upon group performance as a function of task structure. *Journal of Personality and Social Psychology*, 1966, **3**, 238–242.

Shima, H. The relationship between the leader's modes of interpersonal cognition and the performance of the group. *Japanese Psychological Research*, 1968, **10**, 13–30.

Tannenbaum, R., & Schmidt, W. H. How to choose a leadership pattern, *Harvard Business Review*, 1958, **36**, 95–101.

Terman, L. M. A preliminary study of the psychology and pedagogy of leadership. *Pedagogical Seminary*, 1904, **11**, 413–451.

(Received February 26, 1970)

Psychological Bulletin
1971, Vol. 76, No. 2, 149–150

INVARIANCE OF d^* IN MEMORY: RESPONSE TO BANKS

HARLEY A. BERNBACH[1]

Cornell University

A refutation is presented to Banks' argument that the invariance of Type II d' (or d^*, the "corrected-for-guessing" Type II d') with respect to probability correct in recall is an artifact. The replicability of that invariance is discussed, as are its theoretical implications. Finally, the meaning of a criterion in the recall situation is discussed.

Banks (1970) has provided a generally excellent review of the application of the theory of signal detectability to human memory. In particular, he included an extremely clear synopsis of my finite-state decision theory (Bernbach, 1967) for recognition memory and confidence ratings in recall. Unfortunately, Banks performed a significant disservice to his readers in his discussion of the application of the theory of signal detectability to recall. One of the most theoretically important findings in this area, first discovered by Murdock (1966), is that Type II d' (or d^*, the "corrected-for-guessing" Type II d') is invariant with respect to variables affecting memory, such as serial position or trial number (Bernbach, 1967). Banks (1970) claimed incorrectly that Murdock's result was merely "an artifact of his measuring procedure [p. 89]."

Three points need to be made, concerning (*a*) the relationship between Type II d' and the proportions from which it is derived, (*b*) the replicability of Murdock's finding, and (*c*) the meaning of "criterion" in this context.

First, Banks noted that hit frequencies and correct-rejection frequencies were virtually complementary, and that that was as though subjects, because responses were forced, placed guesses in the lower categories of confidence. (The confidence-rating method for determining Type II d' is commonly used, and it was used for all the instances discussed in this paper.) But how are any of these low-confidence responses correct, as many are, when the guessing probability is virtually zero? In my studies using consonant trigrams as responses, the probability of a correct response by guessing is approximately 1/8000! It is more likely that Banks simply observed that hits and false alarms tend to covary; the theory of signal detectability tells us whether or not that covarying is all criterial. (Clearly, it is not in recognition memory.) Thus, Murdock's finding may have been not an artifact, but rather a coincidence.

If the problem is reduced to the possibility of a coincidence, the record can be set straight by looking at replications of the invariance of d^* with respect to variables, such as serial position or number of presentations, that affect the probability of correct recall. At the time Banks wrote his article, two replications involving serial position and three regarding number of presentations (Bernbach, 1967) had been reported. The invariance has since been replicated by Bernbach[2] and by Bernbach and Bower (1970). An important point is that what replicates is the invariance of d^*, not of Type II d'. A correction for guessing derived from the theory that predicts the invariance (as discussed in Bernbach, 1967, p. 476) was necessary to obtain the invariances shown in my Figure 14 (Bernbach, 1967). It would be a robust artifact indeed that would fit into a theory in that way!

The third point is perhaps most important as regards Banks' conclusions. The criterion in question here is not some point above which a response will be made; rather, it is a point above which some particular confidence rating will be made. (In fairness to Banks, it should be pointed out that this same confusion pervades the paper by Murdock, 1966.) Criteria can affect the proportion of correct responses only if subjects are permitted to censor low-confidence responses. There is no criterion problem in short-term memory, *as long as subjects are forced to make some response to every item*. This means that Banks' argument against forced recall should be completely reversed. What remains may be a problem of correcting for guessing (as was done by Bernbach, 1967, to transform Type II d' to d^* where appropriate). This is not a problem to be solved by the application of the theory of signal detectability, however; it is a problem solvable only by some specific theory of memory or of the memory trace.

---

[1] Requests for reprints should be sent to Harley A. Bernbach, Department of Psychology, Cornell University, Ithaca, New York 14850.

[2] Bernbach, H. A. Recognition and recall in short-term memory. Paper presented at the meeting of the Psychonomic Society, San Antonio, Texas, November 1970.

REFERENCES

BANKS, W. P. Signal detection theory and human memory. *Psychological Bulletin*, 1970, **74**, 81–99.

BERNBACH, H. A. Decision processes in memory. *Psychological Review*, 1967, **74**, 462–480.

BERNBACH, H. A., & BOWER, G. H. Confidence ratings in continuous paired-associate learning. *Psychonomic Science*, 1970, **21**, 252–253.

MURDOCK, B. B., JR. The criterion problem in short-term memory. *Journal of Experimental Psychology*, 1966, **72**, 317–324.

(Received October 26, 1970)

Manuscripts Accepted for Publication in the *Psychological Bulletin*

Perception and Discrimination as a Function of Stimulus Orientation: The "Oblique Effect" in Man and Animals: Stuart Appelle*: Department of Psychology, State University of New York at Brockport, Brockport, New York 14420.

Maintenance Schedules and Hunger Drive: An Examination of the Rat Literature: Roy B. Weinstock*: Department of Psychology, Martha Washington College, Fredericksburg, Virginia 22401.

---

* Asterisk indicates author for whom the address is supplied.

Psychological Bulletin
1971, Vol. 76, No. 2, 151–152

CONFIDENCE-RATED RECALL, d_r, AND TESTS OF BERNBACH'S FINITE-STATE THEORY IN RECALL

WILLIAM P. BANKS[1]

Pomona College

The measure d_r, a signal detection index of discriminability between correct responses and intrusions in recall, is discussed, and some questions about its meaning and validity are raised. One of possibly many methods for testing Bernbach's finite-state theory in recall without d^* (i.e., d_r) is suggested.

In a recognition test the subject is shown some new and some old items (distractors and list items, respectively), and indicates which are new and which are old. Since the subject must use memory information to distinguish old and new items, a detection measure (A_g, d', or d_s; see Banks, 1970) of his ability to discriminate old and new items is properly a measure of his memory for the old items. While uncontaminated by criterial effects, the detection measure will reflect such factors as the similarity between old and new items in addition to memory. One of the most attractive features of Bernbach's (1967) finite-state theory of memory is that it takes explicit account of the nonmemory factors in recognition d', as expressed in the following equation (p. 468, Equation 6):

$$d' = (1 - \delta)d^*, \qquad [1]$$

where d' is the empirical measure of recognition performance, and δ and d^* are theoretical quantities, δ being the probability of forgetting an old item and d^* being a function of the discriminability between old and new items that exists when matching list items to items in memory.

An efficient method for obtaining a detection measure from recognition is to have subjects rate each item on the recognition list with their confidence that it is old, and subject the results to a Type II analysis (Banks, 1970, pp. 82–86) to obtain an A_g or d_s. Sometimes the Type II technique is used in recall, and here its outcome has a different meaning from that found in recognition. In recognition the old and new items are provided by the experimenter, but in recall the subject provides all the old (recalled) items and new items (intrusions) himself and then rates them. The d' yielded by a Type II analysis of recall is best conceived of as a measure of the subject's ability to discriminate his *intrusions from his correct responses.* By analogy with Pollack's (1959a, 1959b) reasoning in treating message-reception data, the Type II measure of recognition should be called d_s (discriminability of stimuli) and the Type II measure of recall, d_r (discriminability of responses). The d^* which Bernbach (1971) says is constant in recall is clearly d_r. When, in fact, detection analyses of recall are made which do not use a Type II technique, d' covaries with amount recalled (Murdock, 1968).

Because d_r depends on the subject to serve up both the signal and noise distributions for analysis, I have expressed some doubts about the measure. In particular, with forced recall (forced list-length matching) and a large d_r there is a good chance that d_r will remain artifactually constant, or nearly so, as a result of guessing habits and the fact that opportunities for intrusions are inversely correlated with amount recalled (Banks, 1970, pp. 89–90). This artifact, if it exists, is specific to the response task, not to the manner in which items are learned or forgotten. Bernbach's various replications of the constancy of d_r all use the same response task and therefore neither augment nor diminish the plausibility of the artifact. Furthermore, the a priori probability of low confidence guesses has little to do with the artifact. The seeming 1/8000 probability of a guessed consonant trigram being correct (Bernbach, 1971) is actually 1/20 or less if the subject gave it a low confidence rating because he was unsure of one of the consonants.

---

[1] Requests for reprints should be sent to William P. Banks, Department of Psychology, Pomona College, Claremont, California 91711.

The question of whether or not to force recall is but one issue in a very complicated problem. (I did not simply argue against forced recall as Bernbach seems to think.) This problem has two aspects: the nature of the noise distribution in recall, and the specification of the "recall criterion" used by the subject. The intrusions given by the subject in recall form the empirical noise distribution, and d_r measures the distance between the mean of the intrusions and the recalled items on the likelihood axis. The value of d_r cannot be interpreted as a measure of strength or discriminability of the recalled items unless the nature of the intrusions is assumed constant at all levels of retention. If the subject produces intrusions more similar to the list items as he learns the list better, he will slide both distributions up the likelihood axis, keep d_r fairly constant, and reduce measured criterion in the process (rated recall usually shows a lowered criterion as level of recall increases; see Bernbach, 1967; Bernbach & Bower, 1970; Murdock, 1966). While some questions about the noise distribution might be resolved by, for example, using the intrusions appearing in recall at various levels of retention as the new items in a recognition test given to another group of subjects who learned the same material, the problem posed by the recall criterion is not so easily attacked. The theoretical distributions underlying d_r depend importantly on the criterion the subject is assumed to use in deciding whether to emit a response (Clarke, Birdsall, & Tanner, 1959). Even forcing recall will not ensure equivalent response criteria across conditions because as recall increases more of the list-length matching can be taken up with high-confidence responses and more low-confidence ones can be edited. Although it may seem that some ingenious solution to the recall criterion problem could be found, equating response criteria is not solely a *practical* problem. To postulate a recall criterion dividing strong, recallable items from weak, nonrecallable ones requires a unidimensional strength theory of recall which many investigators would not want to assume. In the context of such a theory, d_r could be a valuable measure, but it requires too many assumptions to be a purely descriptive index of performance.

I have suggested that any conclusions drawn from rated recall should be supported by additional evidence (Banks, 1970, pp. 90, 97). Bernbach's theory is not tied to d_r, and in view of the present doubts about the meaning of d_r it seems he would be wise to turn to tests of different predictions of his theory in recall. An obvious prediction, not put to a test as far as I know, is contained in Equation 1. If percentage (forced) recall estimates $(1 - \delta)$, then the recognition d' for one group of subjects should rise linearly with the recall of another group in the same learning task. The 1967 statement of Bernbach's theory seems to predict, in addition, that the slope of this linear function is d^* (i.e., d_r), but this prediction would probably have to be modified to include a variable expressing the similarity between old and new items on the recognition list.

REFERENCES

Banks, W. P. Signal detection theory and human memory. *Psychological Bulletin*, 1970, **74**, 81–99.

Bernbach, H. A. Decision processes in memory. *Psychological Review*, 1967, **74**, 462–480.

Bernbach, H. A. The invariance of d^* in memory: Response to Banks. *Psychological Bulletin*, 1971, **76**, 149–150.

Bernbach, H. A., & Bower, G. H. Confidence ratings in continuous paired-associate learning. *Psychonomic Science*, 1970, **21**, 252–253.

Clarke, F. R., Birdsall, T. G., & Tanner, W. P., Jr. Two types of ROC curves and definitions of parameters. *Journal of the Acoustical Society of America*, 1959, **31**, 629–630.

Murdock, B. B., Jr. The criterion problem in short-term memory. *Journal of Experimental Psychology*, 1966, **72**, 317–324.

Murdock, B. B., Jr. Serial order effects in short-term memory. *Journal of Experimental Psychology*, 1968, **76**, (4, Pt. 2) 1–15.

Pollack, I. On indices of signal and response discriminability. *Journal of the Acoustical Society of America*, 1959, **31**, 1031. (a)

Pollack, I. Message repetition and message reception. *Journal of the Acoustical Society of America*, 1959, **31**, 1509–1515. (b)

(Received November 10, 1970)

VOL. 76, NO. 3 SEPTEMBER 1971

Psychological Bulletin

FROM PROBABILITY LEARNING TO SEQUENTIAL PROCESSING: A CRITICAL REVIEW

MARI RIESS JONES[1]

Ohio State University

Two-choice probability learning has been the classic situation in which to test formulations of stimulus-sampling theory. Failures of several assumptions of stimulus-sampling models in binary prediction studies are detailed in this review, and an attempt is made to document the direction of current research. Recent trends are examined in research bearing on sequential variables, memory, instructions, discrimination, and age. A number of studies in these areas suggest, contrary to the no-memory, conditioning assumptions of several statistical models, that subjects approach the prediction task in an active, hypothesis-testing manner, often relying on prior events in the sequence. Current thinking that attempts to describe these aspects of behavior, often in a cognitive or information-processing framework, is discussed.

In reviewing over a decade of research in probability learning, Estes concluded in 1964: "There seems to be little room for doubt that it is possible to abstract from the phenomena of probability learning evidence for the pervasive operation of a relatively simple form of conditioning, or associative learning . . . [p. 121]." The form of associative learning Estes proposed was based on the now familiar assumptions of Estes' pattern model: (*a*) the two-choice situation can be represented by an assembly of N mutually exclusive patterns of stimulation each of which becomes connected to predictive responses on an all-or-none basis; and (*b*) exactly one of the patterns is sampled by the subject on each trial with probability $1/N$; (*c*) the probability that the pattern sampled on a trial becomes conditioned (if not already so) to the response reinforced on that trial is a constant, c, over trials.

Estes' interpretation was immediately challenged by Anderson (1964), who noted that statistical models that assume random sampling of stimulus elements cannot handle the overwhelming evidence that subjects respond to prior events in a binary sequence. Citing evidence from schedules in which the event on trial n is conditional on the outcome of trial $n - 1$ (Anderson, 1960; Engler, 1958; Hake & Hyman, 1953), Anderson (1964) pointed out, "If subjects have memory traces of past events when they receive a conditional schedule, then it can hardly be doubted that they have memory traces of past events when they receive a frequency schedule [p. 132]."

To a large extent the evidence since this controversy has favored Anderson's viewpoint, and the area has witnessed the decline of simple associative no-memory statistical models such as the pattern model and its non-Markovian precursor, the linear model (Estes & Suppes, 1959). Theoretical attention has been redirected to the role of memory and the subjects' use of sequential as well as other discriminative cues in predicting completely random (frequency) schedules and various nonrandom event schedules.

The purpose of the present review is to trace the development of research in sequential learning from the probabilistic orientations of the early 1960s to the current concentration on sequential processing. Emphasis is pri-

---

[1] The author wishes to express her gratitude to James R. Erickson and Jerome L. Myers for their constructive and insightful criticisms of an earlier draft of this manuscript. Requests for reprints should be sent to Mari R. Jones, Department of Psychology, Ohio State University, 1945 North High Street, Columbus, Ohio 43210.

© 1971 by the American Psychological Association, Inc.

marily on empirical developments as they relate to emerging theories of sequential prediction. The majority of this empirical work has focused on the two-choice prediction paradigm in which a subject is required to predict with response $A_i (i = 1,2)$ one of two events $(E_i; i = 1,2)$ on each of a series of trials. The review, therefore, limits itself to studies in this two-choice context, except where variations may offer helpful interpretations.

Theories that have addressed themselves to sequential prediction paradigms are of two broad types: cognitive and stimulus-response conditioning models. Few models in either category have sufficient flexibility to span the entire range of data that is reviewed in this article. For this reason, the relevant models are briefly discussed in their respective empirical contexts in order to provide a framework for the research. Detailed theoretical expositions, however, are beyond the scope of this article.

Organization of the review has followed the direction of research in sequential prediction since 1960. Because several reviews (e.g., Estes, 1962, 1964; Weir, 1964) of early work in probability learning exist, redundant discussion of many works was avoided here except when mentioned briefly for appropriate historical continuity. The topics that are considered fall into the following categories: sequential variables, memory, instructions, discrimination, and age. A large body of research relating probability learning to motivation and incentive is omitted; the nature of this research suggests a separate treatment of it.

SEQUENTIAL VARIABLES

Frequency Schedules

Early work in probability learning (Grant, Hake, & Hornseth, 1951; Humphreys, 1939) presented subjects with frequency schedules in which two events were randomized over the entire series of trials with the sole restriction that the more frequent of the two occur a certain proportion of the time (π). In 1953, Hake and Hyman introduced conditional event schedules in which the probability of an event on trial n was dependent on the event occurrence of the previous trial. More recently, the control of higher-order conditional properties (e.g., runs of events) has become common. Accordingly, this section considers research on these different schedules separately.

Frequency theories. Explanations for predictive performance with schedules in which cues associated with prior events have the same correlation with subsequent reinforcing events are numerous (Anderson, 1960; Bruner, Goodnow, & Austin, 1956; Estes, 1959; Estes & Suppes, 1959; Feldman & Hanna, 1966; Gambino & Myers, 1967; Lordahl, 1970; Restle, 1961) and often difficult to test directly with frequency schedules. Estes' statistical models are an exception in that they are addressed primarily to random schedules. The linear statistical model assumes that $P(A_1)$ is a linear function of the response probability on the previous trial:

$$P(A_{1,n+1}) = (1 - \theta)P(A_{1,n}) + \theta\pi,$$

where θ is a learning-rate parameter. The pattern model, described earlier, predicts a similar learning curve:

$$P(A_{1,n+1}) = \left(1 - \frac{c}{N}\right)P(A_{1,n}) + \frac{c}{N}\pi.$$

Both models predict $P(A_1)$ asymptotes at π. Predictions of the two differ with respect to asymptotic response variance and sequential statistics that reflect the probability of a response after different response-event combinations, $P(A_i|E_jA_k)$, or after event runs of length h, $P(A_i|hE_j)$; $i, j, k = 1, 2$; h varies from 0 to the total number of consecutive E_j events. Both models predict that the probability of an A_i response should increase over successive occurrences of the E_i event (i.e., positive recency), but the pattern model fairs somewhat better in fits to certain sequential predictions (Friedman, Burke, Cole, Keller, Millward, & Estes, 1964; Suppes & Atkinson, 1960).

Restle's (1961) schema model assumes that subjects base their predictions on the saliency and frequency of occurrence of homogeneous runs of length, h. The model can predict negative recency and, in some circumstances, positive recency. Gambino and Myers (1967) also suggest that subjects respond to run cues, but subjects are assumed to generalize predictions of a given run length according to a bidirectional gradient to other runs along the run-length continuum. This model, among others (e.g., Lordahl, 1970), has not been directly tested with frequency schedules.

Several theories assume that subjects respond primarily to the sequence of immediately prior events (Anderson, 1964; Feldman, 1963; Feldman & Hanna, 1966; Friedman, Carterette, & Anderson, 1968; Lordahl, 1970). Anderson's model assumes a subject integrates information from prior trials according to a weighted averaging process in which the weights are independent of outcomes of previous trials (i.e., path independence). Theories of Lordahl and Feldman and Hannah both attempt to define subsets of events that the subject uses to test hypotheses. Assessing the subject's verbal descriptions of binary patterns (length eight), Lordahl determined 39 unique hypotheses which placed different emphases on the subject's perceptions of runs, symmetry, and repeating patterns and so could be classified into seven categories. Feldman and Hanna (1966), on the other hand, assume that subjects generate and test hypotheses about the event sequence on each trial, retaining those that result in successful prediction. Subjects are assumed to respond to blocks of prior outcomes as discriminative states defined by conditional probabilities with subsequent events. A modified Foulkes procedure (1959) permits assessment of states up to lengths of five, as well as predictions about the order in which states are discriminated.

More general, and less precise than the foregoing theories, are the problem-oriented formulations of Bruner et al. (1956). Subjects are viewed as engaging in a sequence of decisions (i.e., a strategy) in order to discover valid cues in the probabilistic prediction task. Three objectives control the subjects' strategies: (*a*) maximizing information gain from each prediction, (*b*) minimizing memory load or cognitive strain, (*c*) minimizing risk of failure (p. 234). Given the difficulty posed by a particular event schedule, the subject may shift from searching for valid cues among runs to using relative event frequency as a cue to problem solution in a frequency schedule.

Frequency research. Frequency schedule data present no clear-cut support for any one theory, but do offer difficulties for a few. Because this research centered about testing predictions of linear and pattern stimulus-sampling models, the variables controlled have related to theoretical questions about the role of reinforcement, independence of path, and stimulus variability. For this reason, variables of concern to an information-processing approach have not been manipulated, and the role of memory and cognitive capacity can only be inferred.

The matching of $P(A_1)$ to π and shifts of $P(A_1)$ with π in random schedules have been interpreted as support for statistical learning models (Estes, 1964), as has the finding that first-order conditional statistics rank order with $P(A_1)$ highest following a correct E_1 prediction and lowest after a correct E_2 prediction (Friedman et al., 1964; Suppes & Atkinson, 1960). However, the assumption of the invariance of θ with π has proved inappropriate for the linear model (Estes & Straughan, 1954), and the discrepancy in estimates of the learning parameter from marginal and conditional statistics has proved difficult for the pattern model (Atkinson & Estes, 1963; Suppes & Atkinson, 1960, Chapter 10). Suppes and Atkinson, using $\pi = .60$, found that the learning parameter estimated from asymptotic first-order conditional data produced too rapidly accelerating marginal learning. This suggests either that learning rate is not constant over trials or that subjects may be learning two separate problems, one based on simple event frequency, and another based on sequential properties.

Further, the positive recency predicted by statistical models has been long known to be a sometimes phenomenon. Negative recency first documented by Jarvik (1951) and subsequently by others (e.g., Anderson & Whalen, 1960; Nicks, 1959; Suppes & Atkinson, 1960) is difficult for statistical models. The existence of a "the other event is due" phenomenon (i.e., the gambler's fallacy) with successive reinforcements indicates that subjects are responding to more than the immediately preceding event. Attempting to show that negative recency is a result of preexperimental response biases, Friedman et al. (1964) presented subjects with 288 trials of a $\pi = .80$ schedule (randomized over 48-trial blocks) after over 800 trials of a variable π frequency schedule. The resulting asymptotic positive recency, often cited as support for statistical models, was assumed to reflect the true state of learning in subjects after removal of the masking effects of preexperimental biases with extended training. Others (Edwards, 1961b; Feldman, 1959; Lindman & Edwards, 1961) have also reported positive recency. However, lending support to Anderson's (1964) comment that different frequency schedules may simply train in appropriate recency behaviors is a study by Jones and Myers (1966). Using two π values (.60 and .75), they compared predictions of sequences in which π was randomized over 20 trials with those in which π was randomized over 300 trials. Randomization over larger blocks resulted in higher mean and variance of run lengths, which in turn produced higher $P(A_1)$ values and more positive recency.

Using a lower π value (.50), Friedman et al. (1968) also found little support for positive recency predictions of statistical models even with extended training (1,500 trials per week for 5 weeks). Sequential responding patterns were heterogeneous, but all subjects were strongly affected by the immediately preceding trial with more remote trials acquiring sizable influence in the fifth week. Anderson's weighted-average model provided data fits superior to those of stimulus-sampling models or the Foulkes analysis recommended by Feldman and Hanna. Superiority of the weighted-average model, which assumes that subjects respond to saliencies of prior events in an additive fashion, reinforces the hypothesis that subjects base responding on several events in short-term memory. However, the model's path-independence assumptions were questionable because the subjects showed differential responding to certain subsequences (i.e., alternations) suggesting that groups of events may be coded together in short-term memory. And wide individual-difference weightings derived for these subjects indicate that coding schemes in short-term memory are not uniform with the highly uncertain $\pi = .50$ schedule.

Recent work bearing on the subject's encoding of random events is offered by Yellott (1969). Yellott compared the N-element pattern model with a modified two-parameter version of the simple linear model under two conditions of noncontingent success reinforcement ($\delta = .80$; 1.00, where δ is the probability that the event on any trial agrees with the prediction of that trial). Several standard frequency schedules ($\pi = .2, .5, .8$) preceded the noncontingent success schedules for each subject. Because the pattern model assumes conditioning occurs only on incorrect trials, while linear models assume the probability of conditioning does not change as a function of prior sampling (unless θ is very large with infinite population), Yellott anticipated that sequential statistics under different noncontingent success schedules should discriminate between the two types of models. Alternation response proportions after a sequence of successes did not decrease as predicted by the linear formulations, but remained constant with both $\delta = .80$ and 1.00, thus supporting the pattern model. Unexpectedly, the N-element model nicely predicted the learning curve for noncontingent success ($\delta = .80$) trials with parameters estimated from sequential statistics, a finding not common with noncontingent event schedules (Atkinson & Estes, 1963). However, with $\delta = 1.00$, the observation of response dependencies ("superstitious solutions") could not be handled by either model. Because the dependencies typically consisted of patterns of event runs, periodic and aperiodic, it appears that subjects attempt to encode frequency schedules in terms of such subsequences.

Related to the finding that the subject may respond selectively to prior event patterns is Anderson's (1966) attack on the linear model's description of the stimulus as a fluctuating population of elements. In a series of experiments, Anderson extended an earlier study by Burke, Estes, and Hellyer (1954) in which some or all of a group of signal lamps initiated each of a series of prediction trials. In an attempt to externalize the sampling process, Anderson postulated that the learning rate on a frequency schedule should increase with the proportion of stimulus elements (signal lamps) activated on each trial. He tested the prediction that the difference between

$$P(A_{1,n+1}|E_{1,n}) - P(A_{1,n+1}|E_{2,n}) = k\theta,$$

where k, the number of signal lamps lit on a given trial, was varied (1, 2, 3, . . . 8) as a within-subjects' variable. Using several values of π (.50, .65, .75) and methods of grouping lamps (variable, or in constant groups of three or four), Anderson found little reliable support for the predicted linear relationship between k and the dependent variable, $P(A_{1,n+1}|E_{1,n}) - P(A_{1,n+1}|E_{2,n})$; differences reported were usually between constant and variable signal conditions, with the latter groups exhibiting smaller learning rates; thus, the concept of the stimulus cue as an unordered population of single elements appears inappropriate.

But the nature and complexity of binary patterns to which the subject responds in frequency schedules seem to render simple descriptions impossible. This is emphasized in research by Lordahl (1970) in which subjects extrapolated two events from randomly determined binary patterns of length eight. Lordahl concluded that as many as 40 differentially weighted hypotheses may be necessary to explain the subject's pattern-oriented approach to a prediction task. But a direct application of this analysis to probability learning paradigms (where trials are paced, events presented successively, and subsequences repeated) would ignore limitations of memory and effects of reinforcement schedule on the subject's capacity to generate complex hypotheses. The study does reinforce the notion that the subject codes binary sequences into subsets of events; the complexity of the code may vary with task parameters.

Finally, research on the perceptual aspects of probability learning casts doubt on the stimulus-sampling assumption that a subject samples a stimulus element or pattern on each trial and that this is conditioned in an all-or-none fashion on some trial. Reber and Millward (1968) presented subjects with frequency schedules ($\pi = .80$) in which no prediction responses were required for 100, 500, or 2,000 fast-observation trials (2 trials per second). In 500 subsequent standard probability learning trials (4 seconds per trial), the observation groups did not differ significantly from appropriate control groups in asymptotic response proportions or in run statistics. But in 5 of the 16 possible statistics conditional on outcomes of four previous trials, observation groups did differ significantly; unfortunately, the authors do not describe the direction the differences took, despite the fact that they occurred after relatively salient patterns (e.g., double alternations). However, the general comparability of results in the fast-observation situation with those in the standard probability learning situation lends support to the Reber and Millward contention that there is a cognitive component in sequential learning that the conditioning assumptions of statistical learning theory ignore.

In summary, extended training does not unequivocally result in the positive recency predicted by statistical learning models. Recency effects are a function not only of initial expectancies but also of the particular frequency schedule, with high π values and randomization over large blocks of trials encouraging positive recency. With lower π values, extended training may eliminate negative recency responding for some subjects, but tendencies to rely on prior groups of events are apparent in individual strategies—findings also not predictable by either pattern or linear models.

Research on frequency schedules suggests that even in this complex situation (*a*) subjects can remember groups of events both initially and after extended training; (*b*) subjects code the events into patterns of runs and alternations, although with frequency schedules the particular solution on which the subject seems to stabilize may vary widely among individual subjects; (*c*) subjects do not sample and condition from a population of unordered stimulus elements on each trial; rather, learning has a large perceptual component suggesting that if active sampling occurs it must be a rapid covert cognitive process.

First-Order Conditional Schedules

Conditional theories. With the exception of the simple linear model and the pattern model, most of the models described in the previous section can be discussed in the context of constrained schedules. However, an extension of the linear statistical model for discriminative situations (Burke & Estes, 1957) replaces simple statistical models presented earlier. Burke and Estes' linear trace model assumes that one set of elements or traces is available for sampling following E_1 events, and a second set is available after E_2 occurrences, and these traces enter into the conditioned stimulus on any trial. The theory predicts that repetition responding, $P(\mathrm{A}_{i,n+1}|\mathrm{E}_{i,n})$, on conditional schedules should approximate (plus or minus correction terms) the conditional stimulus specifications, $P(\mathrm{E}_{i,n+1}|\mathrm{E}_{i,n})$.

First-order conditional research. Since Hake and Hyman (1953), one common means of varying the statistical structure of binary sequences has been manipulation of first-order event dependencies, $P(\mathrm{E}_{i,n+1}|\mathrm{E}_{i,n})$, usually holding marginal event frequencies constant. Engler (1958) varied $P(\mathrm{E}_{i,n+1}|\mathrm{E}_{i,n})$ as a within-subjects' variable (.25, .70, .75, .90) over 120-trial blocks with π shifting after 240 trials (.75, .25), and as a between-subjects' variable (values ranging from .10 to $.96\frac{2}{3}$) with π constant (.75 or .50). Asymptotic conditional responding was not a function of π, but tended to follow and exceed $P(\mathrm{E}_{i,n+1}|\mathrm{E}_{i,n})$, a finding consistent with Burke and Estes' predictions. Sequential dependencies affected marginal learning rate so that when π changed, learning was faster if the dependency structure remained unchanged.

The studies of Hake and Hyman (1953) and Engler (1958) were followed by Anderson's (1960) parametric evaluation of conditional schedules with $\pi = .50$ and $P(\mathrm{E}_{i,n+1}|\mathrm{E}_{i,n})$ varied from .1 to 1.00 in .1 increments for 300 trials. $P(\mathrm{A}_{i,n+1}|\mathrm{E}_{i,n})$ again followed first-order event properties exceeding the objective values particularly at higher conditional values. Neither the Burke and Estes model nor a model that assumes the subjects condition either repetition or alternation responses could describe the subjects' tendencies to alternate more than the objective event sequence with lower conditional values. The study not only demonstrates that persistent recency effects observed in frequency schedules are a function of sequence structure, but it also shows that subjects respond to event alternations as readily as to event runs.

Transfer performance of subjects presented with these conditional schedules (300 training trials) revealed that when shifted to a $P(\mathrm{E}_{i,n+1}|\mathrm{E}_{i,n}) = .50$ schedule, subjects change repetition responding in the direction of the transfer schedule but do not approach the asymptote of subjects trained on this completely random schedule. Only groups receiving initial schedules with $P(\mathrm{E}_{i,n+1}|\mathrm{E}_{i,n}) = 1.00$ or 0.00 demonstrated rapid response shifts that approximated response proportions of subjects receiving $P(\mathrm{E}_{i,n+1}|\mathrm{E}_{i,n}) = .50$. This is compatible either with the notion that subjects trained on partial reinforcement schedules are less able to discriminate schedule shifts or are more resistant to extinction than subjects trained on continuous schedules.

Other transfer studies (Goodnow & Pettigrew, 1955; Goodnow, Rubinstein, & Lubin, 1960) have used first-order conditional variations to manipulate run structure. The data suggest that similarities of training and transfer sequences in terms of run lengths determine the direction of transfer. This finding is relevant to Witte's (1964) study in which $P(\mathrm{E}_{i,n+1}|\mathrm{E}_{i,n})$ was .65, .50, or .35 for three groups in a training schedule with $\pi = .50$ for 300 trials. Over 900 transfer trials, where $P(\mathrm{E}_{i,n+1}|\mathrm{E}_{i,n})$ was .50 for all groups, repetition responding exceeded .50, a finding consistent with the Burke and Estes (1957) trace theory but not with Restle's (1961) model. Neither model could handle the run statistics in transfer. In fact, the transfer data revealed long-lasting effects of early training; subjects who saw many long runs in training still showed positive recency over 900 transfer trials. Witte's randomization of π over blocks of 300 trials in transfer may have encouraged such a result. Such randomization would produce more long runs of both events than a sequence in which marginal frequency was more constrained, thus making training and transfer sequences relatively less discriminable in run structures for subjects with $P(\mathrm{E}_{i,n+1}|\mathrm{E}_{i,n}) = .65$ in training and so sustaining the early response patterns of these subjects.

To summarize, subjects' repetition responding is roughly related to first-order sequential properties in that $P(\mathrm{A}_{i,n+1}|\mathrm{E}_{i,n})$ usually follows but exceeds $P(\mathrm{E}_{i,n+1}|\mathrm{E}_{i,n})$. But this dependent variable does not completely reflect the subjects' behavior, since analysis of run curves suggests that subjects are not responding simply to previous events but also to subsequences determined by conditional restraints, and these subsequences play an important role in mediating transfer. It is for this reason that the trace-conditioning theory has proved inadequate, finally. But Restle's assumption that subjects rely on accurate memories of run lengths is also inadequate. It slightly overpredicts repetition responding for data of Engler (1958) and Goodnow et al. (1960), but grossly mispredicts repetition responding and recency curves in Witte's long-term study.

Unfortunately, these conditional analyses offer primarily negative evidence for conditioning theories, while offering few clues to promising alternative formulations. Schedules designed to test assumptions of conditioning theories may have so obscured other psychological processes that the experimenter as well as the subject must guess what is going on. Processes such as hypothesis testing of run structure, if they exist, probably operate at an extreme handicap in these relatively uncertain schedules. The solution, however, is not to assume that no memory or no hypothesis testing exists in these complicated circumstances, but to control such processes first in less complex sequences.

Higher-Order Conditional Schedules

Attempts to isolate the relative importance of various event subsequences in probabilistic predictive tasks have led to more direct control of higher-order sequential properties. Relevant theories fall into three categories: (*a*) simple learning models that assume the subjects' condition either to prior event sequences or to certain homogeneous subsequences (Gambino & Myers, 1967; Vitz & Todd, 1967); (*b*) simple hypothesis-testing or cognitive-processing models (Bruner et al., 1956; Feldman & Hanna, 1966; Myers, 1970; Restle, 1961); (*c*) hierarchical models that assume the subject learns or perceives certain subsequences initially and then progresses to more general forms of the organization of the sequence (Restle, 1967; Simon & Kotovsky, 1963; Vitz & Todd, 1969; Wolin, Weichel, Terebinski, & Hansford, 1965). In one way or another, all these models have made relatively general assumptions about sequential prediction, but direct comparisons of the models are few because appropriate tests are often specific to particular kinds of sequential constraints. All assume that the subjects respond in some way to groups or subsequences of events. That subjects respond to subsequences as units is demonstrated in a study of Ludvigson (1966) who examined resistance to extinction as a function of pattern length (1, 3, 7 successive homogeneous events) and number of pattern repetitions (3, 5, 7 repetitions). Number of repetitions of the pattern unit, and not pattern length, increased resistance to extinction.

Direct manipulation of event runs in a prediction task resulted in sequences ($\pi = .75$) composed primarily of run lengths less than 5, lengths between 5 and 11, and lengths greater than 5 (Derks, 1963). Derks confirmed that negative recency effects are specific to sequences composed primarily of short runs and that gambler's fallacy responding disappears with experience with long runs. This finding indicates that the processes determining recency responding under various types of schedules are similar; the degree of control over such processes is less with random and first-order conditional schedules. Control of the mean and variance of the distribution of runs has further clarified recency effects. Gambino and Myers (1966) combined high (7.5) and low (4.5) mean run distributions factorially with three levels of run variance (low, medium, high). Increasing mean run length and variability increased repetition responding and positive recency.

This behavior seems consistent with run theories such as Restle's (1961), but in 1966 Restle revised his schema model as a result of a direct test. He constructed two sequences ($\pi = .50$) with run lengths of two and five only, with $P(E_i|E_iE_i)$ either .50 or .286 in one of two experiments. In a second experiment, he compared four run-length combinations: (1-4), (2-3), (2-8), (6-9). In both experiments, responding at uncertainty points (e.g., after two consecutive events in the first experiment) was found to be a function of the sum of the run lengths, not the ratio as his 1961 theory implies. Restle suggested instead that subjects either discriminate run lengths and match to the conditional probability of the next event at uncertainty points or they fall back on first-order conditional information and simply use the preceding event. Such an explanation can account neither for the overshooting of first-order conditionals (Anderson, 1960; Engler, 1958; Gambino & Myers, 1966; Witte, 1964) nor for the observed imperfect responding at certainty points (Gambino & Myers, 1966; Restle, 1966).

Gambino and Myers (1967) tackled these problems and assumed subjects inaccurately discriminate run lengths so that run termination or continuation generalizes to other run lengths. Increases in stimulus uncertainty, a function of run-length variance, results in decreases in the subjects' discriminative ability. The model nicely describes both Restle's data (1966) and other data resulting from variations in mean and variance of runs (Gambino & Myers, 1966). Increased variability resulted in increased repetition responding and the likelihood that subjects would overestimate longer run lengths (i.e., commit perseverative errors). Support for the theory was found in the behavior of the generalization parameter, γ, which increased with increases in mean run length and run-length variance. The linear conditioning parameter θ, however, revealed that some run-length combinations are more readily learned than others, a finding not predicted by the theory.

In a further test of the model, Myers, Butler, and Olson (1969) constructed sequences of runs with lengths 4 and 5 or 1 and 5 in which the proportion (β) of longer runs was either .2, .5, or .8 ($\pi = .50$). The model described the subject's relative performance on the different run combinations, breaking down primarily with sequences having many single alternations. But the estimated learning-rate parameter again failed to be invariant across conditions, and γ did not vary as expected with the mean and variance of the run distribution. Typically, errors at certainty points were a function simply of the number of run continuations between the two run terminations with the 1-5 sequence producing more errors than the 4-5 sequence. Repetition responding at uncertainty points was not affected by run variance, but decreased systematically with decrements in the proportion of long runs.

Two related questions confront the run-length models: Is the assumption that runs are the primary discriminative unit correct? What types of relationships among these units are learnable and how are they learned?

Regarding the first question, several studies already reviewed testify to the potency of run cues. Another demonstration is offered by Halpern, Lantz, and Schwartz (1969) in which run cues were pitted against tonal cues as event predictors. Auditory cues were perfectly correlated with event occurrences, uncertainty being introduced by variations in discriminability of the two cues (Δ intensity $= 0, 5, 9$ deviations in decibels [sound pressure level] from a standard of 70 decibels) for six different run-length conditions (randomized sequences and structured sequences varying in mean and variance of run length with $\pi = .50$). The presence of run cues decreased prediction accuracy in the structured sequences when auditory cues were highly discriminable. If sequences were random (randomization procedure unspecified), the subjects relied on discriminable auditory cues and were more accurate than subjects experiencing structured sequences at points of uncertainty in the sequence. It appears that subjects learn to attend to run cues with structured sequences and cannot switch attention at points in the sequence where run cues are unreliable.

But to what extent can subjects attend to reliable sequential cues other than runs? Strub and Erickson (1968) and Erickson and Cuchural (1969) examined subjects' ability to use second-order conditional information. They controlled marginal and first-order event contingencies (.50) using two levels of second-order conditionals ($P[E_{i,n}|E_{j,n-2}] = .92$ or $.72$; $i, j = 1, 2$). In both studies, increased accuracy reflected the subjects' use of the higher-level contingency, but with low-contingency levels subjects tended to fall back on the previous event evidencing repetition responding. The fact that the run structure covaries with second-order conditionals makes it difficult, as the authors pointed out, to determine to what extent the subjects are indeed responding to a single cue, two trials back, or to resulting run patterns.

Wolin et al. (1965) presented three groups of subjects with binary sequences generated by different sources such that the sequences contained subsequences of simple patterns. Rules that insured accurate predictions could be derived from the various subsequences; these were of the form "never an E_2 after an E_2" "always an odd number of E_1s between two E_2s," and so on. Subjects tended to reproduce sample subsequences, responding primarily to homogeneous runs rather than statistical rules. However, there was some evidence that subjects were responding to rules that determined run lengths (i.e., switch on odd trials, continue on even trials). A second experiment introduced a parallel between binary patterns (AB and AAB) and "words," with the same three rules generating partially random "word" contingencies thus serving as substitutes for grammatical rules. Contingencies defining words were rapidly learned, as were some simple rules. With more difficult rules, subjects responded by matching "word" frequencies, but not accurately anticipating the "words." In no case could the subjects verbalize rules.

Other complex contingencies studied include periodic functions in which the probability of an event varies as a function of trial number (Young, 1967) and complex conditional schedules spanning several trials (Feldman & Hanna, 1966). Feldman and Hanna generated a complex conditional sequence using a modified Foulkes procedure. Subjects' predictions over 1,000 self-paced trials indicated that they respond to nonrun as well as run contingencies defined by conditional properties as discriminative states by the model. But the order in which subjects learn the states, which is some function of state length, was not completely handled by the authors' descriptive model.

None of these studies directly confronts the problem of determining the primary discriminative unit, if one exists, used by the subject. They suggest that while subsequences varied in difficulty for subjects, the reinforcement schedule itself strongly determines how the subject used different subsequences to predict. Other paradigms involving immediate recall may have greater potential for an answer to this.

If we assume with Gambino and Myers, among others (Restle, 1961, 1967; Vitz & Todd, 1967, 1969), that as a first approximation, runs are the primary discriminative or coding unit, then the assumption that subjects do not respond to patterns of runs is a handicap for a theory striving for generality.

Butler, Myers, and Myers (1969) varied the proportion of times a run of a given length was repeated successively ($\eta = .25$, .50, .75) in three sequences differing in run components (2-3, 4-6, 2-8). Run redundancy (η) significantly affected repetition responding at uncertainty points such that with 4-6 and 2-8 combinations, repetition responding was minimal for the $\eta = .75$ condition and maximal for $\eta = .50$ condition, suggesting that these subjects were responding to runs of runs. As with the Myers et al. study (1969), perseverative errors did not decrease with decrements in run variance; η did not affect perseverative errors, while run-length combination did. Groups receiving the 2-8 sequences made more perseverative errors than those receiving the 4-6 sequences, who in turn, were less accurate with the longest run length than the 2-3 groups. This, as the authors pointed out, suggests that variability as defined by the difference between the two run lengths determines perseverative error responding.

Limitations of such a model are put in clearer light by a study of Jones and Erickson (1968) in which mean and variance of run structure were identical for two groups, but one group received a recurrent pattern of run lengths, while the other received randomly arranged run lengths. Contrary to Gambino and Myers' predictions, subjects responded to sequential regularities, particularly when such runs followed the longer run lengths. This finding agrees with earlier work of Rose and Vitz (1966) in which rules applicable to one-half of the trials in sequences ($\pi = .50$) generated different patterns. Patterns of the same length were responded to more accurately if they represented simple modifications of the run structure (e.g., $E_2E_1E_1E_1E_2E_2$) than if they were complex (e.g., $E_1E_1E_2E_1E_2E_2$). In some conditions, subjects learned run-run contingencies, and, often these regularities followed the longer run lengths.

A modification of the run-generalization model that permits the learning of run-run contingencies assumes (*a*) with probability *c*, a subject will learn a particular run-run pairing on a given trial, and, with probability $1 - c$, a subject will retain the wrong run-run hypothesis; and (*b*) generalization occurs only after predictive errors (Myers, 1970). The model predicts that in a 100% noncontingent success schedule, subjects will exhibit only periodic solutions of run lengths.

Myers used a noncontingent-event-training, noncontingent success transfer design to test this prediction. Two levels of η, the probability run repetition, were crossed with two training sequence conditions (2-5, 4-5). Periodic solutions in transfer were not observed; rather, subjects tended to generate solutions based on combinations of remembered run codes. The findings were interpreted in terms of a limited-capacity information-processing model in which higher error rates are attributable to more complex and taxing solutions. The fact that subjects in the 2-5 sequences generated "more complex" solutions in the noncontingent success transfer and also showed higher error rates in training was interpreted as support for this view.

In summary, direct manipulation of subsequences has indicated that (*a*) subjects respond appropriately in varying degrees to the subsequence reinforced with runs of homogeneous events and alternation patterns most readily detected; (*b*) increments in mean run length, number of runs, and run continuations between terminations increase perseverative errors; (*c*) repetition responding at uncertainty points increases with mean run length and reflects the subject's sensitivity to run-run contingencies and redundancies in the sequence; (*d*) given similar lengths of subsequences, subjects learn contingencies based on simple run modifications more readily than complex contingencies; (*e*) subjects rarely verbalize statistical rules that generate run contingencies, but rather describe sequences in terms of run combinations.

Neither the Restle nor the Gambino and Myers theory can handle these data adequately. Restle's theory, because it does not permit a confusion factor, cannot handle the subjects' decreased accuracy at certainty points or systematic changes in repetition responding at uncertainty points. The Gambino and Myers model fairs better, albeit in some cases with inappropriate changes in parameter estimates. But, the Gambino and Myers identification of generalization with variance of run structure has proved of only limited value, for it appears that either the range of run lengths or the number of run continuations between terminations is more critical in determining responding at certainty points. Since range, number of runs, and variance have often been confounded in tests of the model, it is difficult to pinpoint the relative effects of the different variables on error responding. In addition, the identification of the learning-rate parameter with incentives and instructions, and not with run combinations, is troublesome in view of the fact that some run combinations (e.g., 2-3) are disproportionately easier to learn than others.

The data suggest that a confusability factor is important in explaining sequential predictions, but it is more likely some function of the number of different run codes a person must process, rather than of statistical variability of the run lengths. The number of codes a subject may identify can, in turn, be a function of the discriminability of the run components. In sequence conditions characterized by a few continuations between the run lengths (e.g., 4-5), it is likely that, at least initially, a subject may be relatively successful both at certainty and uncertainty points using one code ("it's 4 or 5") to predict both run lengths. That is, the subject does not immediately recognize two distinct run lengths. But in sequences with many continuations between the component run lengths (e.g., 1-5), the subject is more likely to attempt a two-code description of the sequence initially.

Increasing the number of codes a subject must remember or organize thus may affect his ability to attend to the ongoing sequence. In the vein of limited-cognitive-processing theories such as those of Bruner et al. (1956) or Myers (1970), taxing cognitive capacity increases error rate. Myers assumes that the number of run codes a subject must process contributes to hypothesis complexity such that with high complexity the subject is more likely to lose count of the current run and guess about the next event with probability equal to $P(E_{i,n+1}|E_{i,n})$. Such an interpretation reflects the fact that the major effect of run redundancies introduced by increasing the proportion of long runs (β) or run repetitions (η) is on repetition responding at uncertainty points, not on counting accuracy. But Myers' additional suggestion—that sequences with very dissimilar run components also contribute to cognitive strain by encouraging the subject to generate more "complex" hypotheses—raises the problem of gauging the complexity of cognitive hypotheses that are based on the same number of run codes. Are certain hypotheses more complex because they result in higher error rates? Another solution to specification of a relationship between independent variables and cognitive complexity is to assume, as previously suggested, that cognitive complexity is simply a function of the number of codes the subject entertains at a time, and this is determined by the discriminability of component runs and the number of irrelevant hypotheses the subject retains. This approach not only avoids scaling the subjects' reported solutions but also allows the possibility that certain run-run contingencies permit elimination of irrelevant hypotheses and thus reduce cognitive complexity. In deterministic schedules, this potential generality becomes an advantage.

Deterministic Schedules

Clearly subjects can learn simple sequential patterns; this was known long before Schoonard and Restle (1961) demonstrated experimentally that subjects quickly become proficient at double-alternation sequences. Rather, the relevant questions are: (*a*) How does the subject learn the sequence? (*b*) Are the psychological processes that control the subject's performance in deterministic sequences similar to those operating when subjects predict frequency and conditional schedules? Do these situations evoke entirely different processes as Estes (1964) suggests, or can psychologists examine processes relevant to probabilistic predictions in the controlled environment of deterministic schedules? Perhaps, as Bruner et al. (1956) insist, commonalities such as similar criteria exist at a higher level, which systematically unites apparently different strategies.

Of the theories that describe predictions of deterministic schedules, few have espoused a simple conditioning approach using the run stem as a discriminative stimulus for the subsequent response. An exception is the model of Vitz and Todd (1967), which is an extension of Bower's (1961) one-element paired-associate model. While fits of the model to error distributions for different patterns (lengths of 6, 9, or 12 events) were fair, the model's assumptions that all stimulus patterns are equally difficult to learn was inappropriate, for continuations of a given run were more readily learned than alternations—a rather common finding even in nondeterministic schedules. Moreover, the learning rate varied with pattern length, the longer patterns being learned more rapidly.

More common are formulations that assume the subject proceeds in some step-by-step manner to construct a representation of the entire sequential pattern (Reber, 1967; Restle, 1967; Simon & Kotovsky, 1963; Vitz & Todd, 1969). Such models vary in their emphasis on mechanisms such as rule learning. Both Hake and Hyman (1953) and Wolin et al. (1965) have argued that with conditional schedules subjects do not attempt or succeed in learning the statistical rules that govern the sequence, but rather they tend to reproduce subsequences of events in order to anticipate coming events. In deterministic schedules, where rule structures remain constant over subsequence presentations, it is reasonable to expect data bearing on this issue. But too often experiments may be interpreted in support of either position.

Early examples of this are studies by Galanter and Smith (1958) who attempted to show that subjects identify recurrent regularities in sequences and construct a linguistic-type model that is isomorphic with these regularities. Using auditory inputs, Galanter and Smith presented subjects with learnable cycles of binary digits. Cycle complexity (number of different runs), not cycle length, determined difficulty. Subjects reported suddenly recognizing the regularities, constructing a model, and testing it. Randomness, or stimulus interference, is assumed to be a factor that prevents adequate model construction, and Bruner et al. (1959) demonstrated that even two random disruptions in a simple pattern ($E_1E_1E_2$) seriously degraded model construction. Analysis of error distributions within the sequence in light of possible rules would have clarified descriptions of subjects' reports. Such sequential analysis bears on the simpler and perhaps more appropriate question at this stage: How and in which order does the subject proceed to derive rules or construct a model? For if model construction exists, the mechanics of it are disappointingly vague.

A promising line of investigation into perceptual processes involved in model construction comes from Garner and Gottwald (1967). Subjects were trained on a simple ($E_1E_1E_1E_2E_2$) and a complex ($E_1E_1E_2E_1E_2$) repeating pattern. In one condition, subjects began at any point in the pattern; in a second, subjects were required to begin at one of the five different sequence positions, thus yielding five groups predicting different initial versions of the same pattern. Analysis of the subjects' pattern descriptions revealed preferred modes of organizing the events, and these methods relied heavily on the use of runs and alternations as units. More importantly, overall pattern difficulty was a function of the speed with which the particular arrangement disconfirmed the subject's initial expectancies (e.g., double alternation) about the sequence. The simple pattern was easiest to learn when started at the onset of the run of three, one of the two preferred descriptions. Subjects learned the complex pattern most readily if the arrangement was consistent with a single-alternation expectancy or if it resulted in an arrangement which rapidly eliminated this expectancy. Related to this is the finding that subjects were generally most accurate in predicting the event following the longest run. This suggests that disconfirmation of dominant expectancies toward alternations, which is supplied by long runs, is a potent source of information for the subject. Longer runs apparently played an important role in both perceptual organization and performance. An earlier study, using binary tones (Royer & Garner, 1966), demonstrated similar trends.

Multichoice paradigms offer some flexibility in further examining the mechanisms of pattern learning. Restle and Brown (1970) presented subjects with four forms (initial, inverted, transposed, and inverted-transposed) of two six-choice serial patterns. Error profiles reflected the subjects ability to abstract systematic properties of the sequence such as trills (i.e., $E_3E_4E_3$), a finding difficult for conditioning theories based on single-event chaining, serial position, or subsequences of prior events. Significantly, a major source of error was attributable to expectancies the subject acquired about the sequence structure.

The parallel, suggested by Restle and Brown (1970) among others (e.g., Wolin et al., 1965) between learning a sequential pattern and learning to understand a language (see Chomsky, 1968) is a useful one. A danger in its application is the substitution of an analysis of sequence structure for analysis of psychological processes that may be a function of syntactic structure. In this regard, Reber (1967, 1969) has elaborated the perceptual learning approach suggested for frequency schedules (Reber & Millward, 1968) with deterministic schedules. In this context, perceptual learning is of an abstract "grammatical" structure, and according to Reber, is "implicit," representing a primitive process requiring no higher-level intellectual factors. Just as a child may learn to speak and understand a language before he can learn formal grammatical rules that determine his language, a subject may learn to use a sequence in a complex way without evidencing formal rule learning. Using artificial grammars with five-letter vocabularies, Reber (1967) showed that with randomly constructed letter sequences, subjects' learning curves reach an early, stable asymptote, while subjects presented with different "sentences" generated by rules continued to improve. With similar techniques, Reber (1969) also demonstrated that transfer proceeds along lines of grammatical structure primarily and is not vocabulary specific.

Implications of this finding are provocative. If elementary patterns such as runs are combined according to rules, then the resulting regularities of these patterns (i.e., the grammatical structure) may be more critical to the subjects prediction mastery and transfer than the simple patterns themselves. Just as run structure defined by conditional rules mediates transfer in probabilistic sequences, higher-order rules or regularities between run lengths may mediate transfer in deterministic schedules. Thus, the order of learning a sequence would be from higher-order regularities down, with subjects finally fitting in defining aspects of the patterns by counting or similar processes. Cues such as constancy of pattern length or discriminative end-members could signal the end of a chunk and facilitate over all applications of the regularities (Fritzen & Johnson, 1969). This implies that cues or expectancies, which the subject may use initially to identify regularities, signal perceptual units that must finally be analyzed more carefully using learned coding devices, such as counting. Thus, counting may be specific to inputs in a later stage of learning.

Some thinking related to this is offered by Vitz and Todd (1969). They assume that subjects respond primarily to homogeneous groups of events and then proceed to combine adjacent units (in the binary case) in a hierarchical fashion. Using Garner's (1962) terminology, complexity of binary (or trinary) patterns can be gauged by the sum of uncertainties[2] evaluated over all levels of the coding process. A series of studies with repeating binary and trinary patterns (Vitz, 1968; Vitz & Todd, 1969) showed that the complexity measure could be used to predict recall accuracy, complexity judgments, and mean number of descriptive words, independently of response measures. Because the model is nonmetric, assuming that the subjects' first level of responding is perceptual, the model cannot be applied to sequential prediction data. However, if it can be assumed that any grammatically generated sequence is approached first in this manner, and secondly in a metric fashion, then this parameter-free approach is promising. The complexity measure correlates highly with Royer and Garner's (1966) measure of organization difficulty ($r = .93$). The Vitz and Todd technique has the potential to handle the relative contributions of π and first-order contingencies, as well as contingencies among runs, albeit each contribution has equal weight. But because the technique is not metric, it would have difficulty with learning based on contingencies between run components that might be more or less discriminable in terms of lengths, a circumstance common in binary sequences.

Other experiments with deterministic schedules are relevant. Both Keller (see Vitz & Todd, 1967) and Derks and House (1965) found that errors are a function of the number of runs in a sequence, not sequence length—findings compatible with Vitz and Todd's interpretation. Derks and House also reported that given the same number of event runs, homogeneous run sequences (sequences in which runs were of the same or similar length) produced fewer errors per response than heterogeneous sequences with mean run-length constant (4 or 8), a finding also consistent with the complexity measure of Vitz and Todd. Thus, this uncertainty index does reflect the notion, useful in both semirandom and deterministic sequences, that complexity is, in part, a function of the number of different run codes with which a subject must deal.

If, as Vitz and Todd suggest, the run-encoding mechanism is a primitive perceptual one, then the existence of responses in a prediction task could conceivably disrupt encoding and constitute a source of error. Others (Bruner, Wallach & Galanter, 1959; Garner, 1962, p. 287) have argued precisely this, but Reber and Millward (1968) seem to suggest that the overt response has little or no effect on responding, at least with randomly arranged events. Experimental evidence is conflicting. Bruner et al. (1959) found that with a simple repeating pattern ($E_1E_1E_2$) subjects not required to make overt responses performed better than those who were. This was not supported by Garner and Gottwald (1967) with patterns of length 5. Using even longer patterns of runs, Derks and Larson (1969) found that subjects who observed the sequence took significantly fewer trials to learn six binary patterns than subjects who made prediction responses or than subjects who merely counted the events in the runs. These authors suggest that where overt responses interfere, they interfere more with the subjects' organization of run-length orders than with perceptions of run lengths. In this regard the data may also be taken to indicate that subjects do not count events in runs initially, so that forcing subjects to count at this point in learning is just as distracting as making a prediction response, thus retarding initial perceptual learning and, so, learning as a whole. But the failure of Garner and Gottwald to obtain interference remains unexplained.

In contrast to perceptual learning approaches are the models of Simon and Kotovksy (1963), Feldman (1963), and Restle (1967). The differences between the latter theories and perceptual learning approaches are in assumptions about the mechanism of learning. All

---

[2] Specifically,

$$H_{\text{code}} = \sum_{i}^{k} H_i$$

where $H_i = H_{\max} + H_{\text{joint}}$ in Garner's terminology, and k = final level of several levels of code analysis.

three assume that the subjects' approach to learning a pattern is systematic, proceeding in a logical hypothesis-testing manner. For Simon and Kotovsky and for Feldman, subjects store a program; for Restle, subjects retain successfully tested grammatical rules. Only Restle's model has been directly tested with deterministic schedules.

Restle assumes that subjects first learn mandatory rules that integer code the run lengths in an all-or-none manner at appropriate locations in the sequence. In sequences with more than two runs, where for a given run stem there may be more than one appropriate continuation, the subject must rely on prior event cues to learn the correct optional rule. A three-state Markov model (see Polson, Restle, & Polson, 1965) describes optional rule learning in which the subject proceeds from an initial guessing state to a hypothesis-testing state; hypotheses are presumably rules about the way runs combine.

Testing the model, Restle (1967) compared transfer performance of four groups on the test sequence ($E_1E_1E_1E_2E_2E_1E_2E_2E_2$) after 72 training trials on different run patterns defined by some of the same mandatory rules. Direction of transfer using mandatory rule analysis was, in general, as predicted, but all groups showed roughly the same mean error profile in the test sequence. Error distributions for mandatory rule locations were fit nicely by the two-state Markov model using the same two parameter estimates for data from all five locations, but the three-state formulation predicted only two of the four optional locations.

The model offers useful ways of conceiving of sequential learning if parameters can be more clearly linked to meaningful experimental variables. The probability of error in the intermediate state for optional rule learning, for example, might vary as some function of the number and type of mandatory rules (or runs) in a sequence. In addition, the model implies that all mandatory and all optional rules are of equivalent difficulty. The data of Vitz and Todd (1967) and Royer and Garner (1966) invalidate this assumption for mandatory rules, while Restle's own data discredit its application to optional rule learning. Lacking specification of variables that might determine important differences in rule difficulty, or sequence complexity, precise prediction of the subjects' order of rule learning is impossible. But the type of model that permits description of error distributions across sequence positions has a distinct advantage in predictive power over the Vitz and Todd approach in which the entire pattern is simply given a difficulty index.

In summary, predictions of deterministic sequences are not accounted for by simple conditioning assumptions whether based on runs or serial position. Perceptual learning theories find some support in structure-mediated transfer and response-interference studies, but the mechanics of such learning remain vague. More complicated models have combined proposals of hypothesis testing with hierarchical processing assumptions (e.g., Restle, 1967) suggesting several discrete stages of learning and permitting detailed theoretical predictions. Although Restle's evidence and an analysis by Derks and House (1965) support such discrete-stage assumptions, these notions remain essentially hollow until effects of experimental variables such as sequence complexity are tied to learning parameters.

Specification of the role of sequence structure at different learning stages may, in fact, offer a means of combining perceptual learning assumptions with more formal hypothesis-testing axioms. At this point the two approaches seem to place relatively different emphases on sequential properties at various learning stages. Hypothesis-oriented theories (e.g., Myers, 1970; Restle, 1967) maintain that subjects initially acquire an unordered catalog of run codes by means of counting, and that higher-order structural constancies such as run-run contingencies may not affect this lower-order learning (Myers, 1970). According to a perceptual learning approach, the subject attends initially to higher-order structural constancies, and these rules would mediate component learning; counting is specific to later stages of learning. Clearly, some index of levels of sequence structure, its regularities, and its relationship to either hypothesis testing or perceptual learning would be useful.

In this respect, the index of Vitz and Todd (1969) represents a useful beginning for attempts to relate sequence structure to psychological processes. The measure reflects not only the number of run codes contributing to sequence structure, but also other potentially salient structural properties such as marginal event frequency. Some justification for linking the measure with a psychological concept such as cognitive complexity derives from its high correlation with measures of learning and organizational difficulty. Despite some limitations in generality (e.g., the lack of allowance for variation of number of coded runs with metric discriminability), it offers a basis for assuming that sequential complexity increases with the number of different runs present. Thus if the subject tests hypotheses about run combinations, the number of hypotheses the subject must test also increases with sequential complexity. Furthermore, assuming some limitations on the subject's cognitive capacity, sequences that require much hypothesis testing and hypothesis revision would tax the subject's cognitive ability. Clearly, such an analysis would suggest that the highly complex frequency schedules, in which a variety of different run codes are haphazardly presented, place maximum cognitive strain on the subject, while semirandom and deterministic sequences with fewer run codes are less complex and result in fewer hypotheses. Thus, the latter sequences are less likely to result in cognitive strain. The notion that cognitive complexity increases as a function of the number of hypotheses the subject must test and retain appears to emphasize a hypothesis-testing process. However, higher-order structural regularity so critical to perceptual learning theorists is not incompatible with this approach if structure is viewed as a vehicle that permits the subject to rapidly eliminate irrelevant hypotheses about run combinations. Similarly, initial expectancies allow the subject to operate with a few hypotheses about the run patterns, and so they may guide the way in which the subject attends to a sequence. In other words, structural properties and initial expectancies determine hypothesis saliency, and so, in effect, may reduce cognitive strain.

The reduction of cognitive strain then may be accomplished by the subject's selective attention to certain sequential cues as determined either by expectancies or higher-order regularities. Strategies apparent with schedules of different complexities can be understood in terms of mechanisms that permit a subject to avoid or reduce cognitive strain. A subject may weigh certain aspects of the sequence differently than others and so eliminate some hypotheses. Initially, the subject's attention is governed by expectancies or dominant hypotheses, and if sequential properties are congruent with these expectancies, the schedule's properties will be rapidly mastered. In deterministic schedules, rapid elimination of a dominant hypothesis facilitates the subjects' learning of the pattern by permitting the subject to shift attention to potentially useful structural constancies, such as longer run lengths. The pattern itself can serve as a mechanism of attention that provides serial integration of cues and so permits the subject to efficiently eliminate irrelevant hypotheses, a possibility suggested by Neisser (1966, pp. 233–234). The long run lengths, by invalidating certain initial expectancies and guiding attention, may serve as anchors about which the subject can systematically attach other run codes in this integration process. Such a strategy is impossible in nondeterministic sequences where the subject cannot eliminate irrelevant hypotheses systematically by using overall structure, or recurrent elements such as long runs, to reduce cognitive strain. Here, attention focuses on different cues to accomplish this. Initial expectancies are apparent in these schedules, primarily in the form of negative recency and alternation responding. But disconfirmation of dominant hypotheses in these schedules is often neither rapid nor consistent. However, in semirandom schedules, where reinforcements consistently define the run codes, the subject's eventual reliance on the current run length is clear. In frequency schedules, the asymptotic stability of repetition responding and idiosyncratic solutions reflect the subject's focusing on recent events and certain salient run patterns. This analysis suggests that the subject tests only the limited number of hypotheses to which he can attend and that the nature of these hypotheses are determined by expectancies and constancies of the reinforcement schedule. Reinforcement then becomes an instructional agent for an attention process, and the potency of a negative reinforcement may be a function of strength of the disconfirmed hypothesis and rapidity with which it is disconfirmed.

The view of reinforcement as an instructional agent, not a conditioning operator, is clearly more compatible with cognitive processing theories as outlined by Bruner et al. (1956) than it is with early learning theories. Similarly, concepts such as cognitive complexity and attention are difficult for early learning theories that emphasized, not uncertainty or complexity, but the conditioning of equally probable stimulus elements. Restricting early research to frequency schedules permitted the simplifying assumption of equal-samping probabilities of patterns associated with prior events (e.g., Atkinson & Estes, 1963), but it also ran the risk of obscuring psychological processes such as hypothesis testing. The evidence that subjects remember and test certain types of stimulus elements and selectively ignore others has accumulated, but often with the aid of more controlled and less complex sequences. Parameter modifications and changes in the definition of the sampled unit from an undefined element to an undefined pattern have not stemmed the tide of disconcerting facts for a conditioning-oriented approach. From the perspective of the subject's performance over a variety of schedules, it is apparent that a major revision of thinking, which includes encoding assumptions as well as specifications of number and kind of elements the subject encodes, is preliminary to a complete model. Precise relationships between response strategies and the number and nature of stimulus elements remain to be specified, but research that acknowledges the requirement that stimulus specification is important may ultimately prove more useful than that which makes little attempt to link model parameters closely to stimulus properties.

MEMORY

Conditioning models cannot explain memory of events prior to the most recent one, but evidence from the previous section reveals that in many cases subjects do respond, often selectively, to prior sequences of events. Models such as the Gambino and Myers model can handle the effects of prior events on learning by assuming the true run length is the conditioned stimulus on each trial. Errors are the result of generalization of responses to conditioned run length, not the result of failure of some memory mechanism. But the notion of a span of memory is clearly compatible with information-processing theories in which memory stores of limited informational capacity are hypothesized (Atkinson & Shiffrin, 1968; Myers, 1970). In line with this are the run-coding models that assume subjects can effectively increase the capacity of a limited memory-store by coding individual events into run chunks, a notion consistent with Miller's original (1956) chunking hypothesis. Thus, recall studies have the potential of revealing (*a*) how the subject codes event sequences to store in memory, (*b*) how many codes a subject can store, and (*c*) what variables affect storage of these codes. In assuming a memory-store notion, however, some evidence to support this interpretation over a conditioning view must be examined.

Recall Memory

Reber and Millward (1965) set out to test the no-memory assumption of simple conditioning models by requiring subjects to recall an event x (x = from 1 through 9 trials) earlier in a frequency schedule. Correct recall was highest for the first three or four trials, but subjects instructed to recall over five trials back were poorer recalling more recent events and overpredicted π. The authors suggested the possibility that memory strain involved in remembering back five or more trials in a random sequence forced these subjects to a strategy of relying on the more frequent event. This study, the first to contradict no-memory implications of statistical learning models (e.g., Atkinson & Estes, 1963), was replicated with several π values (.1, .5, .6, .7) in 1968 (Millward & Reber, 1968). A more detailed analysis revealed that recall decrement with remote events could be attributed to the number of runs in the $n-x$ span. Memory was limited to four runs, and when the subject did not recall the remote event, his guessing probability was determined by π.

Probing the subject's memory of a frequent schedule, Vitz and Hazan (1969) confirmed that subjects do chunk in these situations. Using π = .70, and three widely spaced probe-recall trials, subjects reported between one and five discrete subsequences of events. Unfortunately, an analysis parallel to that of Millward and Reber on determinants of recalled chunks was lacking. Use of a random schedule and one π value clearly limits the manipulative power the experimenter has over determinants of the subject's recall.

A recent study by Myers (1970) brings greater control to the sequence. Using sequences with two run lengths separated by either one or three continuations (2-3, 5-6; and 2-5, 4-7) or with four run lengths separated by one continuation (2-3-4-5; 4-5-6-7), Myers tested subjects' recall of the last five run lengths in a series of 10 probe-recall trials. Recall accuracy of runs decreased as the distance between the runs increased and as the number of run lengths increased. This finding parallels the effects of these variables on perseverative errors in prediction paradigms and provides additional support for Myers' hypothesis that both number of run lengths and the distance between the two contribute to hypothesis complexity.

While the effect of run-run contingencies on recall has not been systematically studied, a comparison of the subjects' relative ability to retain a sequentially presented series of symbols ordered either randomly or with a pattern was presented by Ross (1969). Subjects retained four to five items with randomly ordered sequences, but significantly more with ordered sequences. Thus, arrangement of codes either increases memory capacity by supplying a hierarchical coding scheme for input, or facilitates retrieval of codes from memory.

In this respect, two studies of recall of binary patterns are helpful. Glanzer and Clark (1963) presented subjects with 256 arrangements of eight binary shapes (.5-second exposure) and recorded reproduction accuracy during a 16-second response interval. They found that accuracy decreased as number of runs increased up to a point where the array approximated a single-alternation pattern. A simple run hypothesis cannot handle the curvilinearity resulting from the subjects' recalling single alternations. Schwartz and Lordahl[3] compared recall of 25 patterns of length 8 and 25 patterns of length 10. In agreement with Glanzer and Clark, they found that the sequences recalled better were characterized by a few runs; in addition, sequences with imbalanced runs (i.e., very long and very short runs) were easier to recall. These findings do not completely agree with Myers' analysis of recall in semirandom sequences in which both the number of runs and the distance between the runs increased recall errors. Recall of the more discriminable run codes in deterministic schedules was superior to that of homogeneous run sequences. Explanation of this discrepancy undoubtedly rests with the relative role of run-run contingencies in the two kinds of sequences. In both, the existence of several distinctly different run lengths may stimulate generation of more run-arrangement hypotheses (Myers, 1970) than several similar run codes. But, the constancy of run-run contingencies in deterministic schedules allows the subject to rapidly eliminate the irrelevant run-based hypotheses in these sequences.

The fact that ordering of runs in a deterministic sequence can affect both how the subject proceeds to learn the sequence and the difficulty of predicting that sequence indicates that there may be more and less efficient orders of encoding the component run lengths. And, correlations between organizational difficulty of a pattern and its recall difficulty (Vitz & Todd, 1969) suggest that encoding order does operate on memory. But specific mechanisms such as role of long run lengths in recall are not clear, often because long runs are confounded with the existence of fewer runs or with run heterogeneity. A technique used by Lachman and Dooling (1968) for comparing free recall of connected discourse with that of randomly ordered words could be useful in this context. Using free recall measures after 2, 4, 6, or 8 presentations of the sequences, they found that organization of the lists reflected in word-order regularities proceeds comparatively slowly with random discourse. They suggested that subjects form small core units around anchor words to which additional units are added on each input trial. Several studies (Garner & Gottwald, 1967; Royer & Garner, 1966) have indicated that subjects do organize at least patterned sequences, and many prediction tasks indicate that perhaps longer run lengths could function as anchor stimuli in these sequences (Garner & Gottwald,

---

[3] Schwartz, B. J., & Lordahl, D. S. The effect of sequence structure on recall. Paper presented at meeting of Psychonomic Science, St. Louis, November 1969.

1967; Jones & Erickson, 1968; Rose & Vitz, 1966; Royer & Garner, 1966).

Memory Aids

Speculations about the role of memory in determining the subjects' strategies raise the question of the usefulness of memory aids that display prior events. In concept identification studies (e.g., Bourne, Goldstein, & Link, 1964), performance increased with increases in the number of available instances up to a limit. Consistent with Posner's (1965) analysis, enhancement from the display was greater with more complex concepts. Few probability learning studies have systematically examined the role of memory aids; many which have displayed prior events (Erickson, O'Hara, & Jones, 1968; Singer & Roby, 1967; Wolin et al., 1965) have not found unequivocal support for the usefulness of such aids. In part, this may be a result of display characteristics and/or trial pacing, but, in part, it may be that the sequences used were too complex to reveal clearly any enhancement of performance. If forgetting plays a role in sequential learning, then a demonstration of this might be clearer in controlled sequences where a correct learning strategy is clearly defined at some points in the sequence.

Using semirandom sequences with runs of 2 and 6 or 5 and 6, Butler (see Myers, 1970) crossed three display conditions (no display; current run display; 12 immediately prior events display) with two instructional conditions (neutral; informed of run lengths). According to Butler, if errors at certainty points in semirandom sequences are the result of the subject's failure to remember the length of the current run at the time of encoding it (i.e., at its termination), then a display of that run should eliminate errors and so equate the normally poor performance on 2-6 sequence with that of the 5-6 sequence. According to the run-conditioning model, a run display group should not differ from the other two display groups. The results revealed an interaction between current run display, sequence condition, and instructions with the 5-6 sequence resulting in more perseverative errors than the 2-6 for those subjects instructed as to what run lengths would be present in the sequence. For subjects uninformed about run structure, performance was, as it usually is, worse on the 2-6 sequence than the 5-6 sequence with the current run displayed. These results suggest that a short-term memory mechanism for current run length undoubtedly operates in predicting semirandom sequences, but that a memory aid cannot compensate for nondiscriminate coding of runs such as might occur with 5-6 sequences where the gap between the two run lengths is small. This study represents the first conclusive evidence favoring a memory-store interpretation of sequential responding over a run-conditioning approach.

In summary, results from both memory displays and probe studies suggest that subjects group event sequences into coded chunks, and that the subject's behavior is more appropriately described by information-processing models than by event or event-run conditioning assumptions. For frequency schedules, the nature of these chunks is not determined, but memory-aid findings point to the likelihood of the subjects' encoding run lengths. Subjects can recall up to three or four immediately preceding events, and four or five chunks out of a long series of randomly scheduled events. The number of possible run codes in a sequence is important in determining recall ability; unless runs are all of the same length, sequences with more runs are more difficult to recall. The relatively greater number of hypotheses or arrangement rules the subject must entertain with complex sequences undoubtedly contributes to the subject's memory impairment in these cases. Work in related areas supports this notion: Denny (1969) found the more complex the problem, the more likely is forgetting. But cognitive complexity may be effectively reduced for a given sequence by the introduction of run-run contingencies that permit the subject to eliminate irrelevant hypotheses, and so improve his recall. Deterministic rules may direct the subject's attention to modes of organizing the run arrangement that enhance recall. The relative difficulty of different arrangements on recall of the arrangement rules themselves and component run codes is not known, but approaches such as Reber's suggest superior recall of rules, and these might mediate transfer of individual components. Other reserach (Dominowski, 1965) supports this expectation.

Finally, forcing the subjects to recall prior events in frequency schedules seems to facilitate an overshooting strategy in which subjects rely heavily on the more frequent event. Reber and Millward suggest that this is the result of memory overload or strain. This is consistent with the view that in the more complex frequency schedules, subjects attend and retain only highly salient cues and base responding on these cues in an uncomplicated manner. In less complex sequences, where subjects recall three or four chunks, the subjects' strategies suggest the subject is attempting to accurately predict both onsets and continuations of runs, but only in deterministic schedules is this problem-solving strategy completely successful.

INSTRUCTIONS

Probability learning theories have suffered from the same handicaps as other theories with respect to instructional variables. Early statistical learning theory presumed "neutral" guessing instructions in which there was ample room for self-instruction. From a theoretical point of view, the study of instructional variables raises the question of what process do instructions affect. Do "neutral" instructions indeed eliminate this process, or do they merely eschew control of it? Instructions clearly have potent effects, yet no model adequately incorporates instructional set into its formal structure so that few proposals have been clearly formulated in instructional research. Rather, instructional effects are often treated not as formal aspects of a theoretical process, but as epiphenomena.

In probabilistic prediction tasks, the paucity of hypotheses regarding instructional effects has resulted in a disproportionate amount of research on an early formulation by Flood (1954). Flood maintained that a subject will approach a mathematically optimal strategy of always choosing the more frequent event if he is convinced that the sequence is truly random. Morse and Runquist (1960), for example, found that subjects creating their own "random" sequence tended to show higher $P(A_1)$ than control subjects. Das (1961) attempted to encourage pure optimal strategies by allowing some subjects to evaluate random sequences periodically, but subjects with supposedly more information about randomness were not markedly more likely to be maximizers. In neither study did a majority of the subjects reveal pure optimal strategy responding. Flood's notion assumes that (*a*) subjects can and do formulate an accurate concept of randomness; and (*b*) in the face of this randomness, subjects will not problem solve but will accept the rationally optimal strategy of maximizing. It is dubious whether the former assumption is valid in view of what is known about subjects' attempts to reproduce random sequences (e.g., Tune, 1965). This seems to leave the second assumption virtually untestable. Despite this, several subsequent studies have persisted in trying to correlate "randomness" instructions with higher $P(A_1)$ levels.

A number of these studies have attempted to define the conditions under which the subject will match response probability with marginal event frequency, as well as conditions that lead to overshooting. Problem-solving instructions (Goodnow, 1955) have been indicated as contributing to matching, but in Goodnow's study a number of variables (monetary incentive, task structure, etc.) covaried with instructional variables, so that it is difficult to draw unequivocal conclusions with respect to instructions.

In 1962, both Nies and McCracken, Osterhout, and Voss varied outcome information and instructional set, respectively, in frequency sequences ($\pi = .70$). In the former, π was randomized over blocks of 50 trials; the most important finding was that subjects who predicted the color of marbles rolling from a box had significantly higher asymptotic $P(A_1)$ than control subjects who predicted binary events without a box. This was interpreted as support for Flood's hypothesis; it was assumed that the box made randomness explicit. However, Nies also reported that no subject maximized completely, and 75 of the 128 experimental subjects reported searching for patterns of runs. McCracken et al. (1962) randomized over blocks of 20 trials in two experiments. They found that subjects instructed to predict events on a trial-by-trial basis matched π and were sensitive to sequential aspects of the schedule (i.e., the evidenced negative recency); subjects instructed to avoid trial-by-trial analysis or to problem solve for blocks of trials exceeded .85 in A_1 proportions and were not sensitive to sequential properties. The study of McCracken et al. represents an improvement in instructional research over earlier attempts to convince the subject of randomness. In this study, instructions were used as a tool to direct the subject's attention to certain aspects of the sequential task. Assuming attention responded to these concrete directives, we can conclude that if subjects attend to the trial-by-trial fluctuations of a sequence, they may be encoding cues such as sequential patterns or runs, whereas if subjects attend to larger blocks of trials, relative event frequency becomes a more salient dimension, and response strategies change accordingly.

These two studies together suggest that subjects tend to expect sequential patterns and when they respond to sequential properties they will not maximize. Peterson and Ulehla (1965) permitted one group of subjects to throw a die for each prediction, while a control group predicted the same prearranged ($\pi = .67$) sequences. In accord with Nies, the experimental group exceeded π, with 6 of the 21 subjects showing a maximizing strategy. Swensson and Beach (1967) challenged their interpretation that the die-throwing set induced subjects to ignore sequential properties and maximize. Using Peterson and Ulehla's data, they demonstrated with run curves of experimental and control groups that the former group was not ignoring run structure. A subsequent study, emphasizing to subjects that they avoid the gambler's fallacy, still was not effective in forcing subjects to abandon run-dependent responding. The authors concluded that subjects who showed a pure maximizing strategy did not do so because "they appreciated the inappropriateness of run dependency [p. 282]." Other studies instructing subjects to ignore sequential properties (Lathrop, 1967) support the finding that subjects cannot readily do this.

In an attempt to directly control the subject's attention to sequential properties rather than eliminate it, Butler (see Myers, 1970) informed one group of subjects of the run lengths and corresponding responses in the sequence and gave no information about run structure to a second group of subjects. The design, outlined in the memory section, also examined three memory-display conditions and two sequence levels. Butler found that either instructions or a memory aid decreased anticipatory errors in both sequences and that the effect was greatest early in learning. However, the instructional effect clearly complemented the memory displays in decreasing perseverative errors for both sequences throughout all trial blocks. This suggests that both an active attention process and perfect memory are necessary for appropriate encoding and use of sequential cues.

In summary, those studies directed toward testing subjects' use of a "randomness" concept in probability have found that, under conditions where a concept of "randomness" must be inferred by the experimenter, a few subjects will maximize, thus pushing $P(A_1)$ above π. But it is not clear that even these subjects are totally ignoring sequential properties such as runs. More important, it is apparent that the bulk of the subjects cannot be forced by instructions into the mold of having latent and/or valid notions their of randomness or of purely optimal strategies. Flood originally noted that subjects may attempt to maximize the number correct, rather than the expectation, when they believe there is some sequential pattern. Instructional studies suggest that a majority of subjects are inclined to believe there are patterns even when randomness is allegedly explicit. This rather potent phenomenon has tended to be remarked upon more often than controlled. In the few studies that have attempted some control of the subject's inclinations, it is clear that instructions can affect how the subject attends to aspects of the prior event sequence, and this activity appears to supplement the subject's memory of prior patterns.

Edwards (1961a) has pointed out that costs and payoffs may be instructions to subjects in probabilistic tasks, but clearly other variables too may encourage self-instruction. The sequence itself represents a powerful instructional agent in the face of "neutral" instructions. This may be particularly true of constrained and deterministic schedules in which "reinforcements" can effectively serve as instructional agents to direct the subjects' attention to certain subsequences of events, and so encourage problem-solving rather than gambling behavior.

Discrimination

Probabilistic discrimination paradigms are characterized by the random presentation of one of several discriminative stimuli prior to each event prediction in a series of trials. Two-choice trials may be distinguished by the cue type (T_i, $i = 1, 2, \ldots R$) presented; the cues are more or less valid indicators of the event outcomes on the respective trial types. The two-cue situation has been most frequently studied and so constitutes the bulk of research reviewed in this section.

Early research was concerned primarily with conditioning theories of discrimination learning (Atkinson, 1958, 1960; Burke & Estes, 1957; Bush & Mosteller, 1951). Burke and Estes, for example, assumed that sets of elements representing the discriminative trials (T_i) became conditioned on each trial to one of the two prediction responses. The probability of an E_1 on T_1 trials (π_1) and on T_2 trials (π_2) represented the likelihoods that elements independently sampled on these trials became connected to the A_1 response. Asymptotically, Burke and Estes predicted

$$P(A_1|T_1) = (1-\omega)\pi_1 + \omega[\beta\pi_1 + (1-\beta)\pi_2],$$

where β is the proportion of T_1 trials, and ω is an index of trial similarity representing the amount of overlap between the two stimulus sets. Some support for the model's predictions that $P(A_i|T_i)$ eventually approximates π_i with low stimulus similarity was reported by Estes and Burke (1955) and Estes, Burke, Atkinson, and Frankmann (1957).

Atkinson's observing response model raises the possibility that the subject ignores the intersection of the two stimulus sets on some trials. Atkinson assumed that if subjects are paying attention on a given trial, they may perceive the stimulus cue as a unique pattern. The model yields 16 Markov states which represent the conditioning status of various stimulus sets to observing responses and prediction responses.

In contrast to simple conditioning models of discrimination learning, recent theories reflect growing interest in the subject's response strategies (e.g., Massaro, 1969). Massaro (1969) described a three-state Markov model in which the subject is conditioned to the appropriate response, to the inappropriate response, or not conditioned. An appropriate response is defined as a response having the greatest likelihood of being correct on a given trial. Massaro's theory, in tacitly assuming that the subject evaluates the strategy most likely to lead to success in a given context, marks a beginning of rethinking in discrimination learning. But, in general, discrimination research has been dominated by simple conditioning models, and the implications of newer notions of response strategies have yet to be developed. Unfortunately, theories such as Massaro's have not offered sufficient guidelines for reorientation of discrimination research which has tended to concentrate on parametric investigations of the effects of three variables. These variables are (*a*) π_2, the probability of an E_1 on T_2 trials; (*b*) cue similarity or ω; and (*c*) β, the proportion of T_1 trials.

A significant amount of the research centered about these variables has not unequivocally supported conditioning interpretations. Just as the complexities of sequential dependencies contributed to the inadequacy of no-memory conditioning models in simple probability learning, discriminative predictions as a function of β, π_i, and ω have cast doubt on conditioning models of discrimination. But in simple probability learning, the decline of conditioning models has been paralleled to some extent by a rise in coding or hypothesis-testing models. This is not the case with discrimination learning. But inadequacies of conditioning models have raised several theoretical issues that may bear on an overall formulation of sequential choice behavior in both cue and noncue paradigms. At present no theory with both the predictive precision and generality of stimulus-sampling theory provides a comprehensive alternative framework for these issues. At least three, not unrelated, questions may be raised: (*a*) Do randomly arranged discriminative cues operate simply as independently sampled stimuli with common elements serving as the only source of confusion? (*b*) Does reinforcement connect stimulus and response elements in a passive trial-by-trial fashion, or does it serve as an informative index to an active strategy-testing subject? (*c*) Is the concept of task complexity relevant in the discrimination paradigm, and, if so, how does it relate to cue frequency and similarity? As the following sections illustrate, research has by no means answered these questions.

Research with π_2

Two conditioning models make different predictions about the behavior of $P(A_1|T_1)$ with π_2. Burke and Estes predict that $P(A_1|T_1)$ decrease linearly with decrements in π_2, while Atkinson's attention model predicts a convex function. A majority of the research with π_2 has tended to favor Atkinson's predictions of the marginal statistics, but neither model fairs well in predicting sequential statistics.

Both Popper and Atkinson (1958) and Atkinson, Bogartz, and Turner (1959), using π_2 values between .15 and .85 ($\pi_1 = .85$), confirmed Atkinson's prediction that $P(A_1|T_1)$ was a U-shaped function of π_2. The $P(A_1|T_2)$ statistic was a decreasing function of π_2 as predicted by both conditioning models.

Myers and Cruse (1968), however, found that the U-shaped function was most apparent when there were relatively few T_1 trials. Both conditioning models were unable to explain certain changes in $P(A_1|T_1)$ with π_2 when the proportion of T_1 trials was varied ($\beta = .25$, .50, .75). Even more troublesome for a simple stimulus-element conditioning approach were the sequential statistics. The

$$P(A_{1,n+1}|T_{j,n+1}T_{k,n}A_{m,n}E_{p,n})$$

($j, k, m, p = 1, 2$) showed that subjects with low π_2 were more likely to repeat a correct response with similar adjacent trials, but more likely to repeat an error if successive trial cues differed. A strategy shift was apparent to a lesser extent in subjects with an ambiguous T_2 cue where similarity in adjacent trials enhanced a general tendency of these subjects to repeat correct predictions. But if $\pi_2 = \pi_1$, subjects did not reveal such shifts asymptotically; instead they perseverated erroneously, regardless of trial type.

The complexity of the subjects conditional responding is difficult for traditional discrimination theories. Massaro's model can predict sequential statistics for the symmetric situation (i.e., $\pi_2 = 1 - \pi_1$) by assuming that reinforcement of the response most likely to be correct (i.e., the appropriate response) on $T_{1,n}$ strengthens the likelihood of an appropriate response on $T_{2,n+1}$, regardless of whether these appropriate responses are A_1 or A_2 predictions. But even this model cannot completely account for predictions where $\pi_1 = \pi_2 = .85$; observed behavior in this situation suggests that a more comprehensive assessment of the role of cues in sequential prediction is necessary. In general, the data suggest that trial types do not serve as neutral stimuli whose elements are automatically conditioned with reinforcement. Rather, different cues seem to signal distinct strategies to a subject, and the failure to find a distinction between the two cues may ultimately result in a simplified error perseveration strategy.

Cue Similarity

Variations in cue similarity determine the extent to which the asymptotic value of $P(A_1|T_1)$ is affected by π_2, according to Burke and Estes. The $P(A_1|T_1)$ should match π_1 asymptotically only when there is no overlap of T_i stimulus elements (i.e., when $\omega = 0$); $P(A_1|T_1)$ should approach the overall mean π [$\pi = \beta\pi_1 + (1 - \beta)\pi_2$] as ω approaches 1.00. Uhl (1964) tested this assumption by manipulating the proportion of common elements (0.0, .50, .75, 1.00) in a matrix of lights. Subjects responded to the sets of lights as distinct patterns, behavior more compatible with Estes' pattern model (Atkinson & Estes, 1963) than with a common elements approach.

Others, using pitch cues, have found some support for the Burke and Estes prediction that $P(A_i|T_i)$ is bounded by π_i with low discriminability (Halpern & Moore, 1967). Massaro, Halpern, and Moore (1968) also varied auditory discriminability ($\Delta I = 0$, 1, 2, 3 decibels) with similar values of π_1 (.90, 1.00) and π_2 (.50, .60) and required half the subjects to identify trial types prior to each prediction. Increasing distinctiveness had the predicted effect of separating the two marginal statistics, although $P(A_1|T_1)$ did not become independent of π_2 with highly distinctive cues. Both studies revealed complex conditional responding; in general, the probability of an A_1 after an A_2 was greater than after an A_1 if two successive cues were identified as different by the subject. If successive trials had the same cue, there were few inversions in the typical rank-order of conditionals.

Finally, Massaro et al. (1968) determined that Atkinson's U-shaped $P(A_1|T_1)$ function was specific to the highly distinctive cue situation. Moreover, increasing the distinctiveness of auditory cues ($\Delta I = 1.5, 3, 6$ decibels) decreased the likelihood that the subject would perseverate a correct A_1 prediction if successive trials were different.

A majority of these findings appear difficult for all but an appropriate response hypothesis, and this theory as formulated by Massaro (1969) has been restricted to the situation where $\pi_1 > .50$ and $\pi_2 < .50$. With this limitation, cue similarity is assumed to increase the likelihood of the subjects leaving an unconditioned state and to decrease the probability of entering the unconditioned state after an inappropriate conditioning. While the model describes conditional inversions reasonably well in the Massaro et al. study for $\pi_1 = .80$, $\pi_2 = .20$, it cannot predict changes in marginal statistics with π_2 under different levels of cue similarity without violating assumptions of parameter independence and invariance.

In summary, neither variations in visual nor auditory stimulus similarity support early reinforcement models of discrimination learning. A definition in terms of number of common elements ignores the fact that subjects may not define elements and overlap as the experimenter does and as Burke and Estes suggest.

The fact that a distinct but ambiguous T_2 affects response strategy on the alternate trial more than a less discriminable T_2 cue indicates that subjects do not respond to dissimilar cues independently over π_i levels. This finding is contrary to the general spirit of a common elements conditioning model where decreases in similarity should result in less, not more, interaction between the two types of trials. The appropriate response hypothesis can handle some, but not all, of the data reflected in the convexity of $P(A_1|T_1)$ with π_2 when the cues are distinctive. Unexplained data suggest that subjects seek and define a distinct strategy for distinctive T_2 trials in the absence of experimenter-defined appropriate strategies.

Proportion of T_1 Trials, β

According to Burke and Estes, increasing the proportion of T_1 trials (β) increases the likelihood of an A_1 response asymptotically:

$$P(A_1|T_1) = \omega(\pi_1 - \pi_2) + (1 - \omega)\pi_1 + \omega\pi_2.$$

Shaffer (1963) in testing the prediction that $P(A_1|T_i)$ increases with β found that only $P(A_1|T_1)$ increased linearly with higher proportions of T_1 trials. The $P(A_1|T_2)$ statistic decreased with increases in β. Others (Summers, 1968) have suggested that the relative contrast of event frequencies on cue trials with overall π may have determined the subjects' responding in Shaffer's study.

A parametric study using three values of β (.75, .50, .25) and three values of π_2 (.85, .50, .15) sheds some light on the relationships to which the subject responds (Myers & Cruse, 1968). While $P(A_1|T_1)$ increased with β as Shaffer reported, $P(A_1|T_2)$ was not significantly affected. Trends in both statistics suggest that the $\pi - \pi_2$ difference resulting from variations in β was not critical in determining the likelihood that subjects would maximize their respective responses to the two cues; the most significant finding was evidence of a complex summation effect of β and π_1 (or alternately $[1 - \beta]$ and π_2) indicating that β emphasizes the relative differences between cue validities, a finding difficult for all current models, including Massaro's.

In summary, several combinations of π_1 and π_2 have, in general, shown that $P(A_1|T_1)$ increases with increases in β. The effect, however, is due primarily to changes in $P(A_1|T_1)$ for $\pi_2 = .50$ conditions, which is markedly lower with low β values. For $\pi_2 \neq .50$, there is a summation effect where cue frequency enhances overshooting of the respective π_i values. These findings together with conditional statistics are difficult for all discrimination models.

Discrimination research has progressed from representations of the stimulating situation as neutral sets of elements to more complex patterns, and from studying the conditioning of a single prediction response to the conditioning of an appropriate strategy. But, problems of isolating the pattern as perceived by a problem-oriented subject, and of pinpointing the strategy deemed appropriate by the subject remain. The relative effects of π_i, β, and cue similarity suggest that subjects are more likely to exhibit distinctive strategies on the two trial types when differences that define them are greatest. If $\pi_1 - \pi_2$ and cue discriminability are great, the subject is more likely to respond with different responses on the two trial types. But, paradoxically for conditioning models, increasing cue distinctiveness can also increase the interaction between strategies when one cue is ambiguous: with complete uncertainty on distinctive but ambiguous T_2 trials, subjects are more likely to be confused about the strategy on T_1 trials. Increasing the proportion of T_1 trials not only decreases this interaction but facilitates subjects' acquisition of distinct strategies when the $\pi_1 - \pi_2$ is high.

The continuum from a subject's point of view appears to be defined by cue validity or uncertainty, not by the proportion of elements that is connected to a prediction response. Reinforcement in this context does not determine accumulated associations of stimulus and response elements, but rather represents an operation that confirms or infirms the likelihood that the subject is attending appropriate cues. Because many of the effects of cue similarity and β on response strategies occurred very early in learning, it is likely that these variables contribute large instructional components that direct the subject into weighing different dimensions of the task as clues to accepting or not accepting certain simplifying response strategies. The relative superiority of Atkinson's observing response model over Burke and Estes' formulations suggests that some type of attentional mechanism that permits the subject to restrict his responding to certain more valid cues may be useful in a nonconditioning context.

This raises the question of the relationship between probabilistic discrimination learning and standard sequential learning. The unifying thread of stimulus-sampling theory has been broken, leaving much of this research in need of justification. Two observations that were apparent in simple sequence learning provide a rudimentary framework for drawing some parallels. Subjects in simple sequential tasks were actively seeking and testing dimensions of the problem. The most salient cues for subjects in that context were usually sequential cues, but if these were too numerous and varied, the resulting cognitive strain forced subjects to simpler strategies based on attending to other highly salient cues. Similarly, in the discrimination task, subjects are encouraged by the nature of the task to seek solutions to an apparently twofold problem. Randomized event frequencies on both trial types make the discrimination task difficult and could rapidly encourage simplified strategies based on event frequency. Subjects must, in effect, solve two simultaneous problems concurrently. Restle and Emmerick (1966) have demonstrated that presenting subjects with several problems simultaneously eliminates recoding in concept attainment tasks. Related to this is Erickson's (1966) finding that subjects presented with several simultaneous probability learning tasks of equivalent π_i values tend to overshoot the π_i values. Thus, the memory or cognitive strain resulting from the subject's attempt to find two distinct solutions to discrimination problems with randomized event schedules could readily force the subject's acceptance of simplified strategies based on marginal event frequencies rather than run cues. Clearly, with regard to the important notion of complexity, this interpretation suggests that the number of problems the subject must simultaneously solve enters into a definition of task complexity.

AGE

Because children's behavior seems less sophisticated than adults', it might be expected that the simple conditioning theories of sequential learning would be appropriate with younger subjects. Children may not be plagued with the gambler's fallacy or with active problem-solving tendencies that characterize adult behavior. But as early as 1958, Ross and Levy found that older children (15-year-olds) do exhibit negative recency in successive predictions with frequency schedules ($\pi = .50, .60$). Atkinson, Sommer, and Sterman (1960) put linear statistical models to a more formal test and found that 10-year-old subjects also display negative recency, contrary to prediction. Tests of the linear model with younger children have not always examined conditional statistics that would reveal negative recency. Jones and Liverant (1960), for example, found that children of 4 to 6 years tended to maximize marginal response probability more than 9- to 11-year-olds, but no information was presented on conditional responding. Observation of nonmatching behavior in itself, however, is difficult for both linear and pattern models of probability learning. In short, developmental studies suggest that even with younger children, simple conditioning models meet with difficulties.

Weir (1964) has reviewed much of the developmental research on sequential choice behavior; to avoid redundancies this section concentrates on the two-choice research since 1964. In general, such research has continued to build a strong case against early statistical formulations primarily on the kinds of evidence already suggested. Alternatively, there has been some support for the analyses advanced by Weir in his 1964 review. Weir assumes that changes in probability learning with age reflect differences in rate of development of complex hypotheses on the one hand and information-processing ability on the other. This discrepancy accounts for the U-shaped relation between age and problem-solving performance where solution is defined as selecting the optimal strategy. Much of Weir's review and analysis applies to the three-choice situation in which a child is reinforced a certain percentage of the time (e.g., 33%, 66%) on one of three knobs and never on the remaining two. But the approach Weir has taken is useful in interpreting some trends in two-choice tasks.

The information-processing approach suggested by Weir ties in nicely with developments in thinking about sequential learning in adults (e.g., Bruner et al., 1956; Myers, 1970). More recently, Bogartz (1966b, 1969) has formalized a limited-capacity trace-storage approach to binary prediction of 4- and 5-year-olds which has much in common with the general information-processing framework. Bogartz assumes that children apply simple rules (e.g., alternation or repetition rules) to either the encoded event trace from the previous trial (if the child was attending), or to the trace of the prior prediction response.

Taken together, these theories raise several questions about two-choice binary prediction: (*a*) Is the U-shaped age function derived from randomly reinforced three-choice situations specific only to those situations? (*b*) Do children use rules, and if so, how does rule application vary with age and other independent variables? (*c*) What other, if any, age-specific processes determine the subject's response to binary frequency and nonrandom schedules?

Frequency Schedules

Goldman and Denny (1963) presented boys (ages 5 to 15 years) with 150 probability learning trials ($\pi = .75$). The median percentage was 86% for the A_1 response asymptotically was 86% for the youngest age group; remaining groups matched π. Craig and Myers (1963) reported more detailed research in which three groups of children varying in age from approximately $5\frac{1}{2}$ to 14 years predicted sequences with $\pi = .60$ or .80 for 200 trials. The youngest group underestimated π, but the older children (ages 9–10 years and 13–14 years) matched π. Moreover, the older children demonstrated negative recency responding while the youngest group tended to alternate responses, particularly with $\pi = .60$. This finding has been replicated by Offenbach (1965) for the $5\frac{1}{2}$-year-olds with π values of .60, .75, or .90, but Offenbach also reported that 9-year-olds underpredicted these π values as well.

The apparent incompatibility of these findings points up the importance of protocol analysis. An excellent study by Derks and Paclisanu (1967) helps resolve such conflicts. Using the most extensive age range to date (3 to 25 years), they presented eight age groups with 200 binary events ($\pi = .75$). Subjects with marginal response probabilities less than .66 (two standard deviations from π) were labeled undershooters; those with $P(A_1)$ greater than .84 were overshooters. Derks and Paclisanu found that a majority of 3- to $4\frac{1}{2}$-year-olds were overshooters and a majority of 5- to 6-year-olds were undershooters. The number of probability matchers increased steadily with age, and the number of subjects demonstrating negative recency increased with age, with the first appearance in the 12-year-old group. Thus, previous studies with slightly different combinations of overshooters would have different group marginal statistics. These authors relate the findings to Weir's notion that very young children fixate on a single response, while slightly older children can organize response output more complexly but are not always able to attend to or process environmental events to synchronize response patterns (such as alternations) with sequence patterns.

The existence of specific response patterns in 3- and 4-year-olds has been dramatically demonstrated by Jeffrey and Cohen (1965) with noncontingent success schedules ($\delta = 1.00$, .50, or .33). In the $\delta = 1.00$ condition, where any response the subject made was reinforced, the 3-year-olds perseverated responses with remarkable homogeneity, and all but one of the 4- to 5-year-old group showed strong response-alternation tendencies. It is interesting to compare these rather simple superstitious solutions with the more heterogeneous ones offered by adults in similar circumstances (cf. Yellott, 1969). Work in related areas (e.g., Hodge & Riege, 1966; Tune, 1965) has amply demonstrated that people behave nonrandomly when discriminable cues to solutions are absent. It appears in the present context that particular expectancies, which vary with age, determine the initial event pattern to which subjects attend. In the noncontingent success schedules, subjects are not forced to revise initial expectancies so their behavior reflects the dominant tendencies of that subject population. The fact that these tendencies interfered more with responding of younger children in the Derks and Paclisanu study suggests that these subjects were not capable of adjusting their responses according to incoming error information when it is presented in random noncontingent schedules. With older children and adults, errors signal the need to adjust the response strategy; younger children reveal not only simpler response tendencies but less sensitivity to errors as cognitive signals.

Nonrandom Schedules

Bogartz (1966b, 1969), noting the biases specific to children, as well as their limited attentional capacities, suggested that traces of the prior response or encoded event, if attended to by the child, will evoke an alternate prediction response (i.e., the alternation rule) in single-alternation sequences. Bogartz further postulated that if the child is not attending to either the event or the response trace he will guess randomly. Assuming a one-trial memory capacity, Bogartz describes a Markov process for each subject determined by two independently estimated parameters. He has demonstrated remarkable success in predicting mean performance accuracy, number of error runs of a given length, and other sequential statistics for the single-alternation performance of a group of 4- and 5-year olds (1966b, 1969), despite variations in individual parameter estimates (1966b).

Implications of the trace model led Bogartz and Pederson (1966) to vary intertrial intervals for 4- and 5-year-olds as a between-subjects variable (3 and 10 seconds). Subjects predicting single alternations with a 10-second intertrial interval were significantly poorer, but still responded with greater-than-chance accuracy. The authors concluded that traces may carry over 10 seconds! Using a within-subjects design and long and short intertrial intervals (total trial times were 16 and 8 seconds, respectively), Bogartz (1969) found again that performance accuracy on single-alternation sequences was poorer with long intertrial intervals. An extension of the trace model that assumed the probability of entering a null trace increased monotonically with an intertrial interval could predict much of these data.

Bogartz's concept of children's "inward-attending" is useful in describing different response-based tendencies that do not always relate to the ongoing event sequence. In extending this concept to other than alternating sequences, Bogartz (1969) assumed that children may use a repetition rule based on the prior event or response trace, as well as an alternation rule, depending on the nature of the sequence or the stage of learning. At this point, the relationship between the subject's selection of a rule and sequence properties is not clearly specified so that it is difficult to predict in advance, except intuitively, what rule the subject will apply in any given sequence. A second difficulty involves the failure to specify the circumstances in which a subject may differentially respond to stimulus traces and response traces, since the possibility that subjects do not always use the same rules for the two types of traces is suggested by Bogartz (p. 386). If rules applicable to response traces reflect age-specific response biases, then to experimentally examine differential rule application in, for example, 4- and 5-year-olds, nonalternating event sequences should present a clearer picture of the relative contributions of stimulus and response traces.

In this regard, Jones[4] compared the relative performance of 3-year-olds and 5-year-olds on alternating [$P(E_i|E_i) = .1$] and repeating [$P(E_i|E_i) = .9$] semirandom sequences and found that both age levels were more dependent on the repeating sequence, despite the fact that it was the nonpreferred sequence for the older children. More importantly, for the 5-year-olds, alternation tendencies interfered with learning a repeating sequence, a finding that suggests that rules specific to response traces or "inward-attention" may interfere with rules applied to the stimulus traces. Thus, failure to attend the stimulus trace may be correlated at some level with regression to age-specific expectancies which represent rules applied to response traces, a correlation suggested by Bogartz (p. 386).

Children's responses to other nonrandom sequences have been reported by Offenbach (1968) and Myers and Myers.[5] Offenbach varied $P(E_2|E_2)$ (0.00 or 0.50) with $\pi = .75$ with three groups of children (7½ to 12 years). All age levels could pick up these contingencies. Myers and Myers used four age levels (mean ages: 5 years, 7 months; 7 years, 8 months; 9 years, 6 months; 11 years, 7 months), presenting each with run lengths of 1 and 5 only or runs of 4 and 5 in which the probability (β) of the longer run was either .5 or .2. At certainty points, errors decreased with age; at uncertainty points, repetition responding decreased with age. All groups showed sensitivity to sequential contingencies, the 4-5 sequence being relatively easier. Even the youngest group exhibited negative recency in the 4-5 sequence with $\beta = .2$. Although the percentage of subjects showing negative recency at age 5 was not reported, the finding suggests that observations of negative recency appearing first at age 12 (Derks & Paclisanu, 1967) are specific to frequency schedules where there are a greater number of different run lengths. Because no 3–4½-year-olds were sampled, the possibility that Weir's U-function is specific only to random schedules remains.

Alternatively, Myers and Myers (see Footnote 5) suggest that such findings are compatible with the Gambino and Myers theory that people respond to runs as imperfectly discriminated cues. Their contention that age differences are attributable to differences in capacity to discriminate among run cues is reasonable, but remains to be formally tested. Moreover, as research on sequential variables has indicated, many data apparently harmonious with run-conditioning models are also consistent with run-encoding assumptions of information-processing models. Experimental distinctions between a discrimination process and run-code confusion may prove difficult.

Thus, in the Myers and Myers study the fact that even 5-year-old subjects showed negative recency with the 4-5 sequence can also be interpreted as support for the notion that children can encode sequential events into run codes if the task is not too complex. Just as increasing the number of run codes in a sequence may tax the cognitive capacity of adults, the greater number of apparent codes in the 1-5 sequence here may have taxed the more limited capacity of children. The relatively better performance of children with the 4-5 sequence supports an earlier suggestion that people may code two similar run lengths as one and thereby reduce the number of codes to be remembered. In the case of children, this simplification permitted them to reveal memory capacity for several prior events. This finding, it should be noted, is also difficult for Bogartz's model which assumes a one-trial trace capacity of memory.

Response Mode

The shift in theoretical emphasis from stimulus-sampling models to information-processing models points up the different conceptions of the role of the response. In the former context, the response had no particular properties and indeed if response tendencies reflected any nonconformities, they could not be handled easily. The nature of response traces, however, becomes relatively interesting in the latter approach in which interfering properties of response traces are crucial.

---

[4] Jones, S. J. The effect of repetition probabilities on children's two-choice learning. Unpublished doctoral dissertation, University of Massachusetts, Amherst, 1968.

[5] Myers, N. A., & Myers, J. L. A developmental study of sequence structure in binary prediction. (Tech. Rep. No. 3), Amherst: University of Massachusetts, 1968.

In two experiments, Bogartz (1965) tested the linear statistical model's ability to handle the sequential responding behavior of very young children (4 years, 3 months to 5 years, 7 months). With frequency schedules having $\pi = .80$ and .50, he found that children using manual predictive responses showed both strong response alternations and position preferences which the linear model could not handle. In the second study, 12 of the 16 subjects in the first experiment were used in a similar test except with verbal prediction responses. Both first-order conditionals and other sequential statistics proved compatible with predictions of the linear statistical model in this experiment. However, in a subsequent reanalysis of the data of Experiment II, Bogartz (1969) found that his trace model with two parameters could describe these data almost as well as the four-parameter linear model, if he assumed children were using a repetition rule.

Since both Kessen and Kessen (1961) and Offenbach (1965) reported alternation behavior with young children using verbal responses, Green and Myers (1968) compared the two response modes using a frequency schedule ($\pi = .60$) and two age levels (mean ages: 4 years, 2 months and 5 years, 8 months). Neither event-alternation nor response-alternation behavior varied as a function of response mode, but response alternation was found to decrease over the 200-trial session. These authors speculated that experienced subjects may have been responsible for Bogartz's superior fit in the second of his two experiments. Asymptotic first-order conditional statistics did not rank order according to predictions of the linear model in either age group, but it is possible with a comparable number of trials there would be greater correspondence with Bogartz' findings.

The interfering role of verbal response traces in the context of an information-trace model has been studied by Bogartz (1969). He presented children with two decks of cards; one deck was a binary single-alternation deck, the second had four different cards, two of which were identical to the members of the single-alternation deck. He required 4- and 5-year-olds to name one or three cards from the interpolation deck after each of 65 predictions of the single-alternation deck. Mean performance accuracy on the alternation sequence decreased from the usual value of .80 to about .66. Bogartz interpreted this as evidence that the naming response enters an interfering response trace into memory and the child applies his alternation rule to this trace. He found that the children's use of the alternation rule was rather resistant to change even with this rather large error rate, a finding that again suggests that children either appropriately or inappropriately do not use errors as cognitive signals to revise strategies.

Transfer

Not surprisingly, investigations on children's ability to respond to shifts in marginal or higher-order event dependencies indicate that sensitivity to sequential shifts increases with age.

Jones and Liverant (1960) found that 9-year-olds were more rapid in learning a π reversal than 4-year-olds, a result that they attribute to the greater experience of older children with reinforcement probabilities. This conclusion seems to ignore any maturational differences as a function of age. Kessen and Kessen (1961) trained two age levels (mean ages: 3 years, 7 months and 4 years, 5 months) for 50 trials with π either .50 or .70 and then shifted half of each group to $\pi = .70$ or .50, respectively, for 50 more trials. The older children tended to shift marginal predictions appropriately, while younger children showed a marked tendency to stick with responding appropriate to the first 50 trials.

Bogartz (1966a) studied the effects of different types of training sequences on the subsequent alternation performance of 4- and 5-year-olds. When training sequences consisted of 12 presentations each of patterns $E_2E_1E_1$ and $E_2E_1E_1E_1$, there was marked interference with performance on the single-alternation sequence. In general, the children's ability to respond accurately to the more complex regularities of the training sequences was poor, but a temporal delay prior to the E_2 in each case enhanced accuracy on the second event (i.e., the first E_1). The effects of such a delay are somewhat difficult for a trace model such as Bogartz's. The delay may have emphasized the respective patterns as units for the children, thus increasing the chances that subjects would recall them together. In any case, it is possible that training subjects to a criterion of learning prior to transfer would have reduced interference effects from the patterned sequences. Variations of first-order event probabilities (.25, .50, .75) for 75 training trials also interfered with alternation performance, but training with simple repeating sequences did not. As in other studies, older children were more accurate in transfer.

Using more extreme first-order conditional cues (.1 and .9) to define alternating and repeating sequences, Jones (see Footnote 4) compared 3- and 5-year-olds over 100 training and 100 test trials in a factorial shift design ($\pi = .50$). Both age levels showed greater sensitivity to sequential properties than in earlier studies, and the event dependency was greater for repeating sequences and for older children. Again, the children revealed limited dependency on the transfer sequence even after 100 trials, with greater interference in transfer occurring in subjects shifted from alternating to repeating than those experiencing the reverse shift.

Despite the limitations of age ranges and types of sequences studied, transfer data reveal that subjects become more sensitive to sequential shifts with age. Existing work suggests that children are most likely to respond to sequential properties early in training, and that sequence shifts late in training are more difficult. The relative nature of the shift is also important: Children are more likely to adjust to alternating sequences after repeating sequences than the reverse shift.

To summarize, in random two-choice schedules there is evidence to support Weir's hypothesis that children's behavior can be viewed as a combination of ability to process external cues and abilities to generate complex hypotheses. Younger children show limited event dependency, and strong response perseveration tendencies, a fact which makes their marginal responding appear optimal. Children of intermediate ages show more complex strategies such as response alternation, and this often results in lower $P(A_1)$ scores. Older children show greater sensitivity to preceding events in a random sequence and evidence negative recency responding.

Applicability of Weir's hypothesis for nonrandom schedules, where different optimal solutions exist, has not been assessed. But in such schedules, all children reveal greater sensitivity to the event structure and to run cues than was apparent with frequency schedules. Evidence of, perhaps, rule generation is suggested in studies using alternating and other nonrandom sequences. Whether this performance reflects children's ability to generate and test formal rules or whether it reflects merely the retention of certain dominant expectancies remains a fine distinction. If children do approach a sequence with the ability to attend to sequential cues, the manner in which this ability changes with age and task complexity becomes important. However, an interpretation of the data suggests that age and complexity may interact.

This interaction can be examined by assuming that at least two important processes change with age: expectancies and cognitive-processing capacity. Age-specific expectancies become apparent in the stereotyped responding characteristic of any situation devoid of discriminable cues. The noncontingent success schedule is the most obvious example of this, but from a child's viewpoint other situations may also seem devoid of identifiable cues. The disproportionate difficulty of frequency schedules may so tax the child's limited cognitive capacity that in this situation he is unable to identify reliable sequential cues or to revise initial expectancies. The fact that 5-year-olds do show sensitivity to run cues with simpler sequences (i.e., few run lengths) supports the notion that the complexity of frequency schedules can contribute to children's apparent insensitivity to sequential properties. A majority of studies have used frequency schedules, and so may have disproportionately taxed children's limited processing capacity, thus demonstrating primarily that in the face of cognitive strain, the subject relies on dominant expectancies. In a related context, Siegel (1969) has shown that there is an interaction of age and task difficulty such that the processing of greater amounts of information is differentially more difficult for younger children.

Thus, strategies that children demonstrate with frequency schedules may reflect a combination of attending to salient nonsequential cues and a reliance on dominant expectancies—both of which may result from cognitive strain. The role of attention, therefore, may be more prominent in children than in adults, because the cognitive capacity of children is more limited. The dominant hypotheses or expectancies on which the subject focuses initially vary in nature and complexity with age and often determine characteristically different initial strategies. In adults, cognitive strain is initially reduced by differential weighting of sequential cues determined by these expectancies. This takes the form of early biases toward alternations of runs and negative recency. The ability to identify sequential cues may be reduced by task complexity. Anderson (1960) reported that adults trained with complex schedules were less sensitive to subsequent sequential regularities than those experiencing simple schedules. In children, frequency schedules may have more marked effects so that initial expectancies often operate on salient nonsequential cues such as recent stimulus or response traces. In young children, response perseveration is congruent with a "correct strategy" in many cases, while response alternation typical of 5 and 6 years is not. Both may persist inappropriately with frequency schedules as a result of children's inability to cope with error information in complex circumstances. The relatively greater prominence of interfering expectancies in children's behavior may be the combined result of attempts to reduce cognitive strain and children's inability to revise hypotheses. Children's greater sensitivity to disconfirmations of expectancies early in training indicates that rapid invalidation of dominant tendencies is a relatively potent source of error information for children as well as adults.

Significantly, in schedules less complex than frequency schedules, some children do reveal focusing on sequential cues. But, effectiveness of structural regularities in permitting children to attend to higher-order sequential cues and so eliminate some irrelevant hypotheses is unknown. If, as some have suggested (e.g., Myers, 1970), structural constraints increase cognitive complexity, children may be less adequate in deterministic sequences than in semirandom ones with comparable numbers of run codes. However, children's ability to use the sequential regularities in language, even at very young ages, as well as their remarkable imitative abilities, would seem to suggest the contrary. Indeed, the fact that children learn to speak before they master the abstraction of counting suggests that higher-order regularities are more readily attended to than component properties. An examination of children's ability to predict and recall deterministic sequences would have significant bearing on theories that emphasize processes such as counting, rule learning, and hypothesis testing in sequential learning.

REFERENCES

ANDERSON, N. H. Effect of first-order conditional probability in a two-choice learning situation. *Journal of Experimental Psychology*, 1960, **59**, 73–93.

ANDERSON, N. H. An evaluation of stimulus sampling theory: Comments on Professor Estes' paper. In A. W. Melton (Ed.), *Categories of human learning*. New York: Academic Press, 1964.

ANDERSON, N. H. Test of a prediction of stimulus sampling theory in probability learning. *Journal of Experimental Psychology*, 1966, **71**, 499–510.

ANDERSON, N. H., & WHALEN, R. E. Likelihood judgments and sequential effects in a two-choice probability learning situation. *Journal of Experimental Psychology*, 1960, **60**, 111–120.

ATKINSON, R. C. A Markov model for discrimination learning. *Psychometrika*, 1958, **23**, 309–322.

ATKINSON, R. C. A theory of stimulus discrimination learning. In K. J. Arrow, S. Karlin, & P. Suppes (Eds.), *First Stanford symposium on mathematical methods in the social sciences*. Stanford, Calif.: Stanford University Press, 1960.

ATKINSON, R. C., BOGARTZ, W. H., & TURNER, R. N. Supplementary report: Discrimination learning with probabilistic reinforcement schedules. *Journal of Experimental Psychology*, 1959, **57**, 349–350.

ATKINSON, R. C., & ESTES, W. K. Stimulus sampling theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology*. Vol. 2. New York: Wiley, 1963.

ATKINSON, R. C., & SHIFFRIN, R. M. Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory*. Vol. 2. New York: Academic Press, 1968.

ATKINSON, R. C., SOMMER, G. R., & STERMAN, M. B. Decision making by children as a function of amount of reinforcement. *Psychological Reports*, 1960, **6**, 299–306.

BOGARTZ, R. S. Sequential dependencies in children's probability learning. *Journal of Experimental Psychology*, 1965, **70**, 365–370.

BOGARTZ, R. S. Variables influencing alternation prediction by pre-school children. I. Previous recurrent, dependent and repetitive sequences. *Journal of Experimental Child Psychology*, 1966, **3**, 40–56. (a)

BOGARTZ, R. S. Test of a theory of predictive behavior in young children. *Psychonomic Science*, 1966, **4**, 433–434. (b)

BOGARTZ, R. S. Short-term memory in binary prediction: Some stochastic information processing models.

In G. H. Bower & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances and theory.* Vol. 3. New York: Academic Press, 1969.

Bogartz, R. S., & Pederson, D. R. Variables influencing alternation prediction by preschool children: II. Redundant cue value and intertrial duration. *Journal of Experimental Child Psychology,* 1966, **4,** 211–216.

Bourne, L. E., Jr., Goldstein, S., & Link, W. E. Concept learning as a function of availability of previously presented information. *Journal of Experimental Psychology,* 1964, **67,** 439–448.

Bower, G. H. Application of a model to paired-associate learning. *Psychometrika,* 1961, **26,** 255–280.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. *A study of thinking.* New York: Wiley, 1956.

Bruner, J. S., Wallach, M. A., & Galanter, E. H. The identification of recurrent regularity. *American Journal of Psychology,* 1959, **72,** 200–209.

Burke, C. J., & Estes, W. K. A component model for stimulus variables in discrimination learning. *Psychometrika,* 1957, **22,** 133–145.

Burke, C. J., Estes, W. K., & Hellyer, S. Rate of verbal conditioning in relation to stimulus variability. *Journal of Experimental Psychology,* 1954, **48,** 153–161.

Bush, R. R., & Mosteller, F. A model for stimulus generalization and discrimination. *Psychological Review,* 1951, **58,** 413–423.

Butler, P. A., Myers, N. A., & Myers, J. L. Contingencies among event runs in binary predictions. *Journal of Experimental Psychology,* 1969, **79,** 424–429.

Chomsky, N. *Syntactic structures.* The Hague: Mouton & Co., 1968.

Craig, A. J., & Myers, J. L. A developmental study of sequential two-choice decision-making. *Child Development,* 1963, **34,** 483–494.

Das, J. P. Mathematical solutions in the acquisition of a verbal CR. *Journal of Experimental Psychology,* 1961, **61,** 376–378.

Denny, N. R. Memory and transformations in concept learning. *Journal of Experimental Psychology,* 1969, **79,** 63–68.

Derks, P. L. Effect of run length on "Gambler's Fallacy." *Journal of Experimental Psychology,* 1963, **65,** 213–214.

Derks, P. L., & House, J. L. Effect of event run structure on prediction of recursive binary sequences. *Psychological Reports,* 1965, **17,** 447–456.

Derks, P. L., & Larson, A. V. Response interference in acquisition of sequential patterns of binary events. *Journal of Verbal Learning and Verbal Behavior,* 1969, **8,** 304–306.

Derks, P. L., & Paclisanu, M. L. Simple strategies in binary prediction by children and adults. *Journal of Experimental Psychology,* 1967, **73,** 278–285.

Dominowski, R. L. Role of memory in concept learning. *Psychological Bulletin,* 1965, **63,** 271–280.

Edwards, W. Costs and payoff are instructions. *Psychological Review,* 1961, **68,** 275–284. (a)

Edwards, W. Probability learning in 1,000 trials. *Journal of Experimental Psychology,* 1961, **62,** 385–394. (b)

Engler, J. Marginal and conditional stimulus and response probabilities in verbal conditioning. *Journal of Experimental Psychology,* 1958, **55,** 303–317.

Erickson, J. R. On learning several simultaneous probability-learning problems. *Journal of Experimental Psychology,* 1966, **72,** 182–189.

Erickson, J. R., & Cuchural, J. Use of second-order dependencies in two-choice learning. *Journal of Experimental Psychology,* 1969, **79,** 575–577.

Erickson, J. R., O'Hara, J. W., & Jones, M. R. Effects of a memory aid in probability learning. *Psychonomic Science,* 1968, **10,** 137–138.

Estes, W. K. Theory of learning with constant, variable, or contingent probabilities of reinforcement. *Psychometrika,* 1957, **22,** 113–132.

Estes, W. K. Component and pattern models with Markovian interpretations. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory.* Stanford, Calif.: Stanford University Press, 1959.

Estes, W. K. Learning theory. *Annual Review of Psychology,* 1962, **13,** 107–144.

Estes, W. K. Probability learning. In A. W. Melton (Ed.), *Categories of human learning.* New York: Academic Press, 1964.

Estes. W. K., & Burke, C. J. Application of a statistical model to simple discrimination learning in human subjects. *Journal of Experimental Psychology,* 1955, **50,** 81–88.

Estes, W. K., Burke, C. J., Atkinson, R. C., & Frankmann, J. P. Probabilistic discrimination learning. *Journal of Experimental Psychology,* 1957, **54,** 233–239.

Estes, W. K., & Straughan, J. H. Analysis of a verbal conditioning situation in terms of statistical learning theory. *Journal of Experimental Psychology,* 1954, **47,** 225–234.

Estes, W. K., & Suppes, P. Foundations of linear models. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory.* Stanford, Calif.: Stanford University Press, 1959.

Feldman, J. On the negative recency hypothesis in the prediction of a series of binary symbols. *American Journal of Psychology,* 1959, **72,** 597–599.

Feldman, J. Simulation of behavior in the binary choice experiment. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought.* New York: McGraw-Hill, 1963.

Feldman, J., & Hanna, J. F. The structure of responses to a sequence of binary events. *Journal of Mathematical Psychology,* 1966, **3,** 371–387.

Flood, M. M. Environmental non-stationarity in a sequential decision-making experiment. In R. M. Thrall, C. H. Coombs, & R. L. Davis (Eds.), *Decision processes.* New York: Wiley, 1954.

Foulkes, J. D. A class of machines which determine the statistical structure of a sequence of characters. *Wescon Convention Record,* 1959, **4,** 66–73.

Friedman, M. P., Burke, C. J., Cole, M., Keller, L., Millward, R. B., & Estes, W. K. Two-choice behavior under extended training with shifting probabilities of reinforcement. In R. C. Atkinson (Ed.), *Studies in mathematical psychology.* Stanford, Calif.: Stanford University Press, 1964.

Friedman, M. P., Carterette, E. C., & Anderson, N. H. Long-term probability learning with a random schedule of reinforcement. *Journal of Experimental Psychology*, 1968, **78**, 442–455.

Fritzen, J., & Johnson, N. F. Definiteness of pattern ending and uniformity of pattern size: Their effects upon learning number sequences. *Journal of Verbal Learning and Verbal Behavior*, 1969, **8**, 575–580.

Galanter, E. H., & Smith, W. H. S. Some experiments on a simple thought problem. *American Journal of Psychology*, 1958, **71**, 359–366.

Gambino, B., & Myers, J. L. The effect of mean and variability of event run length on two-choice learning. *Journal of Experimental Psychology*, 1966, **72**, 904–908.

Gambino, B., & Myers, J. L. Role of event runs in probability learning. *Psychological Review*, 1967, **74**, 410–419.

Garner, W. R. *Uncertainty and structure as psychological concepts*. New York: Wiley, 1962.

Garner, W. R., & Gottwald, R. L. Some perceptual factors in the learning of sequential patterns of binary events. *Journal of Verbal Learning and Verbal Behavior*, 1967, **6**, 582–589.

Glanzer, M., & Clark, W. H. Accuracy of perceptual recall: An analysis of organization. *Journal of Verbal Learning and Verbal Behavior*, 1963, **1**, 289–299.

Goldman, A., & Denny, J. Ontogenesis of choice behavior in probability learning and sequential programs. *Journal of Genetic Psychology*, 1963, **102**, 5–18.

Goodnow, J. J. Determinants of choice-distributions in two-choice situations. *American Journal of Psychology*, 1955, **68**, 106.

Goodnow, J. J., & Pettigrew, T. F. Effect of prior patterns of experience upon strategies and learning sets. *Journal of Experimental Psychology*, 1955, **49**, 381–389.

Goodnow, J. J., Rubinstein, I., & Lubin, R. Response to changing patterns of events. *American Journal of Psychology*, 1960, **73**, 56–67.

Grant, D. A., Hake, H. W., & Hornseth, J. P. Acquisition and extinction of a verbal conditioned response with differing percentages of reinforcement. *Journal of Experimental Psychology*, 1951, **42**, 1–5.

Green, D., & Myers, N. A. Children's motor and verbal responding in a two-choice situation. *Journal of Experimental Psychology*, 1968, **77**, 314–316.

Hake, H. W., & Hyman, R. Perception of the statistical structure of a random series of binary symbols. *Journal of Experimental Psychology*, 1953, **45**, 64–74.

Halpern, J., Lantz, A. E., & Schwartz, J. A. Cue properties of the event run in choice discrimination learning. *Journal of Experimental Psychology*, 1969, **80**, 237–242.

Halpern, J., & Moore, J. W. Two-choice discrimination learning as a function of cue similarity and probability of reinforcement. *Journal of Experimental Psychology*, 1967, **74**, 182–186.

Hodge, M. H., & Riege, W. H. Individual response patterns in binary choice tasks with nondiscriminable stimuli. *Perception and Psychophysics*, 1966, **1**, 421–425.

Humphreys, L. G. Acquisition and extinction of verbal expectations in a situation analogous to conditioning. *Journal of Experimental Psychology*, 1939, **25**, 294–301.

Jarvik, M. E. Probability learning and a negative recency effect in the serial anticipation of alternative symbols. *Journal of Experimental Psychology*, 1951, **41**, 291–297.

Jeffrey, W. E., & Cohen, L. B. Response tendencies of children in a two-choice situation. *Journal of Experimental Child Psychology*, 1965, **2**, 248–254.

Jones, M. H., & Liverant, S. Effects of age differences in choice behavior. *Child Development*, 1960, **31**, 673–680.

Jones, M. R., & Erickson, J. R. Patterns of event runs in probability learning. *Psychonomic Science*, 1968, **13**, 317–318.

Jones, M. R., & Myers, J. L. A comparison of two methods of sequence randomization. *Journal of Experimental Psychology*, 1966, **72**, 909–910.

Kessen, W., & Kessen, M. L. Behavior of young children in a two-choice guessing problem. *Child Development*, 1961, **32**, 779–788.

Lachman, R., & Dooling, D. J. Connected discourse and random strings: Effects of number of inputs on recognition and recall. *Journal of Experimental Psychology*, 1968, **77**, 517–522.

Lathrop, R. G. Perceived variability. *Journal of Experimental Psychology*, 1967, **73**, 498–502.

Lindman, H., & Edwards, W. Supplementary report: Unlearning the gambler's fallacy. *Journal of Experimental Psychology*, 1961, **62**, 630.

Lordahl, D. S. An hypothesis approach to sequential predictions of binary events. *Journal of Mathematical Psychology*, 1970, **7**, 339–361.

Ludvigson, H. W. Response units in the prediction of simple event patterns. *Journal of Experimental Psychology*, 1966, **72**, 355–360.

Massaro, D. W. A three state Markov model for discrimination learning. *Journal of Mathematical Psychology*, 1969, **6**, 62–80.

Massaro, D. W., Halpern, J., & Moore, J. W. Generalization effects in human discrimination learning with overt cue identification. *Journal of Experimental Psychology*, 1968, **77**, 474–482.

McCracken, J., Osterhout, C., & Voss, J. D. Effects of instructions in probability learning. *Journal of Experimental Psychology*, 1962, **64**, 267–271.

Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 1956, **63**, 81–97.

Millward, R. B., & Reber, A. S. Event-recall in probability learning. *Journal of Verbal Learning and Verbal Behavior*, 1968, **7**, 980–989.

Morse, E. B., & Runquist, W. N. Probability matching with an unscheduled random sequence. *American Journal of Psychology*, 1960, **73**, 603–607.

Myers, J. L. Sequential choice behavior. In G. H. Bower & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory*. Vol. 4. New York: Academic Press, 1970.

Myers, J. L., Butler, P., & Olson, D. Run lengths and probabilities in binary prediction. *Journal of Mathematical Psychology*, 1969, **6**, 444–455.

Myers, J. L., & Cruse, D. Two-choice discrimination learning as a function of stimulus and event probabilities. *Journal of Experimental Psychology*, 1968, **77**, 453–459.

Neisser, U. *Cognitive psychology*. New York: Appleton-Century-Crofts, 1966.

Nicks, D. C. Prediction of sequential two-choice decisions from event runs. *Journal of Experimental Psychology*, 1959, **57**, 105–114.

Nies, R. C. Effects of probable outcome information on two-choice learning. *Journal of Experimental Psychology*, 1962, **64**, 430–433.

Offenbach, S. I. Studies of children's probability learning behavior: II. Effect of method and event frequency of two age levels. *Child Development*, 1965, **36**, 951–962.

Offenbach, S. I. Studies of children's probability learning behavior. VI. Effect of event sequences. *Journal of Experimental Child Psychology*, 1968, **6**, 460–469.

Peterson, C. R., & Ulehla, Z. J. Sequential patterns and maximizing. *Journal of Experimental Psychology*, 1965, **69**, 1–4.

Polson, M. C., Restle, F., & Polson, P. G. Association and discrimination in paired-associate learning. *Journal of Experimental Psychology*, 1965, **69**, 47–55.

Popper, J., & Atkinson, R. C. Discrimination learning in a verbal conditioning situation. *Journal of Experimental Psychology*, 1958, **56**, 21–25.

Posner, M. I. Memory and thought in human intellectual performance. *Bristish Journal of Psychology*, 1965, **56**, 197–215.

Reber, A. S. Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, 1967, **6**, 855–863.

Reber, A. S. Transfer of syntactic structure in synthetic languages. *Journal of Experimental Psychology*, 1969, **81**, 115–119.

Reber, A. S., & Millward, R. B. Event memory in probability learning. *Psychonomic Science*, 1965, **3**, 431–432.

Reber, A. S., & Millward, R. B. Event observation in probability learning. *Journal of Experimental Psychology*, 1968, **77**, 317–327.

Restle, F. *Psychology and the judgment of choice. A theoretical essay*. New York: Wiley, 1961.

Restle, F. Run structure and probability learning: Disproof of Restle's model. *Journal of Experimental Psychology*, 1966, **72**, 382–389.

Restle, F. Grammatical analysis of the prediction of binary events. *Journal of Verbal Learning and Verbal Behavior*, 1967, **6**, 17–25.

Restle, F., & Brown, E. R. Serial pattern learning. *Journal of Experimental Psychology*, 1970, **83**, 120–125.

Restle, F., & Emmerick, D. Memory in concept attainment. Effects of giving several problems concurrently. *Journal of Experimental Psychology*, 1966, **71**, 794–799.

Rose, R. M., & Vitz, P. C. The role of runs in probability learning. *Journal of Experimental Psychology*, 1966, **72**, 751–760.

Ross, B. M. Sequential visual memory and the limited magic of the number seven. *Journal of Experimental Psychology*, 1969, **80**, 339–347.

Ross, B. M., & Levy, N. Patterned predictions of chance events by children and adults. *Psychological Reports*, 1958, **4**, 87–124.

Royer, F. L., & Garner, W. R. Response uncertainty and perceptual difficulty of auditory temporal patterns. *Perception and Psychophysics*, 1966, **1**, 41–47.

Schoonard, J. W., & Restle, F. Analysis of double alternation in terms of stimuli and responses. *Journal of Experimental Psychology*, 1961, **61**, 365–367.

Shaffer, J. P. Effect of different stimulus frequencies on discrimination learning with probabilistic reinforcement. *Journal of Experimental Psychology*, 1963, **65**, 265–269.

Siegel, L. S. Concept attainment as a function of amount and form of information. *Journal of Experimental Psychology*, 1969, **81**, 464–468.

Simon, H. A., & Kotovsky, K. Human acquisition of concepts for sequential patterns. *Psychological Review*, 1963, **70**, 534–546.

Singer, E., & Roby, T. B. Measurement of sequential behavior by successive differences. *American Journal of Psychology*, 1967, **80**, 276–281.

Strub, M. H., & Erickson, J. R. Second-order dependency in probability learning. *Journal of Experimental Psychology*, 1968, **78**, 261–268.

Summers, S. A. Alternative bases of choice in probabilistic discrimination. *Journal of Experimental Psychology*, 1968, **76**, 538–543.

Suppes, P., & Atkinson, R. C. *Markov learning models for multiperson interactions*. Stanford, Calif.: Stanford University Press, 1960.

Swensson, R. G., & Beach, L. R. Instructions about randomness and run dependency in two-choice learning. *Journal of Experimental Psychology*, 1967, **75**, 279–282.

Tune, G. S. The nonindependence of responses to a two-choice situation. *Journal of General Psychology*, 1965, **72**, 173–178.

Uhl, C. N. Effect of overlapping cues upon discrimination learning. *Journal of Experimental Psychology*, 1964, **67**, 91–97.

Vitz, P. C. Information, run structure and binary pattern complexity. *Perception and Psychophysics*, 1968, **3**, 275–280.

Vitz, P. C., & Hazan, D. N. Memory during probability learning. *Journal of Experimental Psychology*, 1969, **80**, 52–58.

Vitz, P. C., & Todd, T. C. A model of learning for simple repeating binary patterns. *Journal of Experimental Psychology*, 1967, **75**, 108–117.

Vitz, P. C., & Todd, T. C. A coded element model of the perceptual processing of sequential stimuli. *Psychological Review*, 1969, **76**, 433–449.

Weir, M. W. Developmental changes in problem-solving strategies. *Psychological Review*, 1964, **71**, 473–490.

Witte, R. S. Long-term effects of patterned reward schedules. *Journal of Experimental Psychology*, 1964, **68**, 588–594.

Wolin, B. R., Weichel, R., Terebinski, S. J., & Hansford, E. A. Performance on complexity patterned binary event sequences. *Psychological Monographs*, 1965, **79** (7, Whole No. 600).

Yellott, J. I., Jr. Probability learning with noncontingent success. *Journal of Mathematical Psychology*, 1969, **6**, 541–575.

Young, F. W. Periodic probability learning. *Psychonomic Science*, 1967, **8**, 71–72.

(Received March 23, 1970)

Psychological Bulletin
1971, Vol. 76, No. 3, 186–204

IDENTIFICATION OF CONTENT AND STYLE: A TWO-DIMENSIONAL INTERPRETATION OF ACQUIESCENCE[1]

P. M. BENTLER[2]
University of California

DOUGLAS N. JACKSON
University of Western Ontario

AND SAMUEL MESSICK
Educational Testing Service

The methodology involved in identifying and verifying content and stylistic sources of variance in personality inventories is reviewed. The evidence indicating that response style variance is of negligible importance appears to be far from conclusive. Previous item-reversal studies, which have been considered to yield unambiguous evidence for specific content responding, are shown to be consistent with a style position which differentiates two acquiescence processes. Agreement acquiescence, such as true-saying, appears to be a relatively minor determiner of responses to structured inventories. Acceptance acquiescence, such as subscribing to or denying heterogeneous attributes as self-descriptive, appears to be a major response determiner. These two types of acquiescence are usually confounded in analyses of personality inventories. Recent studies evaluating the existence of acquiescence variance are summarized.

In recent years there has been a vigorous literature concerned with the interpretation of the nature of responses to structured personality inventories. However, there is by no means universal agreement regarding such important issues as the conditions under which specific item content determines response consistency. The impetus of a number of early studies led Cronbach (1950) to draw attention to the fact that achievement test scores were substantially influenced by characteristics other than item content, such as response biases favoring particular alternatives or differential guessing tendencies. Following a number of empirical studies in the personality domain, Jackson and Messick (1958) emphasized the importance of these response biases by highlighting a distinction between *content* and *style.* They concluded that the major common factors in personality inventories of the true–false or agree–disagree type then available, such as the MMPI and the California Psychological Inventory, were "interpretable primarily in terms of style rather than specific item content [p. 247]." These styles were conceptualized as response consistencies to general item characteristics, such as desirability level, direction of wording, and type of format, as distinct from specific item content.

More recently, Rorer (1965) has come to a radically different conclusion, namely, that the existence of response styles is "without any foundation whatsoever [p. 150]." However, it is important to note that Rorer redefined the term "response style" to specifically exclude several major types of stylistic responding previously discussed in the literature, such as desirability and extremity. He used the term "style" to refer only to "a tendency to select some response category a disproportionate amount of time independently of the item content [p. 134]." Thus, Rorer's specialized use of the general term "style" makes his position appear at first glance to be all-embracing as well as extreme:

> Though the list of studies purporting to find evidence of response styles in these inventories is long and varied, it is equivocal and unconvincing. The evidence that content rather than style determines responses is unequivocal and overwhelming [p. 145].

Rorer's extreme position has fostered the in-

---

[1] Portions of this paper were presented at the 1965 annual meeting of the American Psychological Association (Jackson & Messick, 1965). This research was supported in part by Grants MH-08691, MH-02878, and MH-16992 from the National Institute of Mental Health and by Grant 151 from the Ontario Mental Health Foundation. We would like to express special thanks to Lawrence J. Stricker, Fred Damarin, and Martin E. Morf for criticisms of an earlier version of this paper.

[2] Requests for reprints should be sent to P. M. Bentler, Department of Psychology, University of California, Los Angeles, California 90024.

terpretation in many quarters that method variance, in the sense of Campbell and Fiske (1959), is essentially nonexistent, a congenial stance for many workers in the area of personality measurement (e.g., Block, 1965). Recently, following an earlier presentation by Jackson and Messick (1965), a number of research workers have begun to question Rorer's interpretation. Campbell, Siegman, and Rees (1967) and Rundquist (1966), for example, have raised a number of objections to Rorer's interpretation of both personality and attitude data; Peabody (1966) focused his objections upon Rorer's interpretations in the attitude domain.

The general purpose of the present study is to attempt to clarify the role of stylistic variables and of specific item content in determining responses to statements designed to assess personality. One specific goal is to seek further definitional and conceptual clarity regarding "acquiescence" and its relation to "content." There are two reasons for focusing this review upon acquiescence. First, despite the impression given by his flamboyant title, "The Great Response-Style Myth," even Rorer (1965) considered the existence of previously identified response styles such as social desirability (called "sets" by Rorer) to be of such importance that "They may account for a significant portion of the response variance on personality inventories, and their measurement and control are essential if the predictive validity of personality inventories is to be increased [p. 151]." Therefore, a broader review of response styles in general is not necessary at this time. Second, recent experimental evidence indicates that acquiescence is best understood as composed of at least two different processes, rather than only one. Virtually all workers in the area of personality assessment have considered acquiescence as synonymous with the tendency to respond "true" to test items, but after the publication of a number of item-reversal studies in the early 1960s this conceptualization was seen to be inadequate. Instead, it seemed plausible to us that the number of "true" responses might reflect the operation, either separately or conjointly, of two conceptually distinct acquiescence processes. Previous data, which had not permitted the separation of these processes, were therefore seen to be ambiguous on this point. This review aims to explicate these two processes, to differentiate one from the other as well as from content-specific responding, and to describe some of the experimental evidence bearing upon this conceptualization.

Content and Style Positions

Proponents of a pure content position in the content style controversy would contend that all stable individual differences on personality instruments reflect the responses of individuals to the specific semantic content of the stimuli. Proponents of the style position believe that some stable individual differences on personality measures represent consistencies in response to general characteristics of the stimuli other than and in addition to specific item meaning. Social desirability, extremity, and acquiescence have often been considered major contributors to noncontent variance. It should be noted in passing that interpretations of response style have frequently included reference to its *interdependence* with content (Jackson & Messick, 1962b), contrary to Rorer's special definition.

The person who says "true" when asked whether or not he is "friendly" may well be saying "true" for a number of possible reasons. A *strong* version of the pure content position would hold that a person answers "true" if and only if he really is friendly; for example, because he smiles when he meets people or because he is courteous. The strong version of the content position, however, can immediately be seen to be unrealistic since it requires virtually perfect accuracy in self-perception and self-report. A *weak* version of the pure content position would hold that a person answers "true" only if he believes himself to be friendly. Both versions of the content position would require evidence for the convergent and discriminant validity for the particular trait in question. If a person reports that he is friendly on a personality questionnaire but is observed by others to behave in an unfriendly manner, the pure content position must account for this discrepancy. If the discrepancy is attributed to consistent inaccuracies in self-description or to idiosyncratic modes of expression, then the nature of these biases would appear to pose an interpretive dilemma for the pure content position. In contrast, the style position (or, more precisely, the content-style position) holds that a person may say "true" for additional reasons other than the actual or perceived truth value of the specific item content. For example, he may say "true" to make himself look good, as when applying for a job, or because he likes to be agreeable, as when conforming to a group, or because he has a relatively undifferentiated self-concept, as revealed in a general tendency to indiscriminantly endorse many traits as self-descriptive.

The distinction between content and style in personality assessment was highlighted by Jackson and Messick (1958) to emphasize the importance of individual response consistencies that appeared to have a basis, not in the specific personality variable of interest to the scale constructor, but in other expressive consistencies in personality influencing the questionnaire response. Jackson and Messick (1962b) elaborated this distinction as follows:

> variance associated with *content* is considered to refer to response consistencies in certain defined assessment situations which reflect a particular set of broader behavioral tendencies, relatively enduring over time, having as their basis some unitary personality trait, need state, attitudinal or belief disposition, or psychopathological syndrome. The *item content* used to elicit such behavioral predispositions may be developed initially on theoretical or *a priori* grounds, may be obvious or subtle, may be direct or indirect, and may be highly relevant or only slightly relevant to some particular prediction criterion: the initial defining property of content assessment is some form of *response consistency* . . . variance associated with *response style* has reference to expressive consistencies in the behavior or respondents which are relatively enduring over time, with some degree of generality beyond a particular test performance to responses both in other tests and in non-test behavior, and usually reflected in assessment situations by consistencies in response to item characteristics other than specific content [p. 134].

This conception of response style implies a number of important and occasionally controversial conclusions:

1. A style is a potentially measurable personality variable or trait. This simple statement implies that the framework of measurement theory and experimental analysis is relevant to the assessment of stylistic tendencies.

2. A style refers to a behavioral consistency operating across measures of several conceptually distinct content traits. The words "conceptually distinct" are important here, since we do not mean "empirically uncorrelated." In this view, a meaningful style score requires some internal consistency across content domains. A style is thus identified by the existence of a latent variable (usually, but not necessarily, a common factor of linear factor analysis) which is associated with response tendencies in several distinct content areas.

3. Styles may be manifest in several conceptually distinct domains of personality. Stylistic consistencies may be found in the areas of perception, cognition, or motivation, for example, in addition to that of the personality assessment area. There is nothing to prevent a particular style from cropping up in many domains.

4. A style may be manifest in responses to test materials presenting content in various formats or by various methods. Identification and verification of any particular style in various modes of content presentation is, of course, an empirical matter. Various styles and content dimensions may, and usually do, exist simultaneously as components of variance in a total score.

5. A large number of conceptually distinct styles may exist. Some of the styles that have been identified represent individual differences in tendencies to respond: desirably, acquiescently, deviantly, defensively, fakingly, extremely, guessingly, and critically (Messick, 1968).

6. The degree of generality of any particular style is an empirical question. In each case, one might be hopeful for a general factor across several content areas, but empirical data are ultimate arbiters here. Any particular individual difference measure of a given style needs psychometric verification including construct validity data, and the verification of interrelationships among various measures of a particular hypothesized style is subject to all the constraints of the validating process.

7. Since a style is a personality variable, it will ordinarily exhibit correlations, linear or nonlinear, with certain other behavioral consistencies or content traits. The extent of correlation is empirically determined, but

bears upon the construct interpretation of the style. A particular style may have correlates with content scores obtained through the same method of measurement or through different methods of measurement. For example, a style observed in personality inventories may have personality correlates in scores obtained from an inventory or from observer's ratings or from behavioral measures (Jackson & Messick, 1958). This point appears to be controversial, since both Block and Rorer seem to feel that the existence of behavior-rating correlates of the first two MMPI dimensions contradict a response style interpretation of these dimensions. Rorer (1965), for example, stated that findings by Block with the MMPI of "correspondence between the ratings and the responses simultaneously rules out either a social desirability or a faking interpretation of the factors [p. 144]."

8. A style will usually attenuate the discriminant validity of personality scales. For example, the tendency to respond in a desirable direction, if present in several conceptually distinct trait measures, may spuriously raise their average intercorrelation. As a consequence, it may be necessary to use experimental or statistical methods for eliminating the effects of response styles on trait scores in order to obtain relatively uncorrelated measures of independent traits. While Rorer (1965) agreed with this point for those styles that he referred to as "sets," such as social desirability, he disagreed with this conclusion for the style of acquiescence, which he viewed as inconsequential. Block (1965) also considered the role in the MMPI of response styles, as method factors contaminating the measurement of content, to be trivial.

9. Style and content are not mutually exclusive; identifying stylistic variance does not necessarily contraindicate the presence of content. Indeed, it may be possible to develop measures of content traits that are relatively free of the effects of response styles. This point needs separate mention because it is often assumed that a belief in the existence of response styles automatically requires disbelief in the possibility of measuring content traits. This is a position taken only very rarely (Berg, 1959).

IS THERE A CONTENT-STYLE CONTROVERSY?

Several publications appearing at about the same time a few years ago had the combined effect of giving the impression that there really is no more content-style controversy: There is no controversy, because styles do not exist. The most influential of these publications were Rorer's (1965) review and Block's (1965) monograph. A number of item-reversal studies also appeared (Lichtenstein & Bryan, 1965; Rorer & Goldberg, 1965a, 1965b), showing sizable consistencies in response to original and reversed forms of MMPI items. These were interpreted as a final deathblow to response styles, since the subjects' observed consistency was attributed to content, not style. Many psychologists have viewed these publications as the culmination of work in the response-style area, as well as a signal for the termination or extinction of the general concept of response style (e.g., Janis, Mahl, Kagan, & Holt, 1969; Mischel, 1968; Nunnally, 1967). It is often forgotten that the work of Rorer and Block was rapidly challenged in various replies. Bentler (1966), Campbell et al. (1967), Edwards (1967), Jackson (1967b, 1967c), and Jackson and Messick (1965) primarily focused their criticism upon Rorer's conclusions regarding personality scales, while Peabody (1966) and Rundquist (1966) focused their evaluations upon what they considered errors in Rorer's conclusions regarding attitude questionnaires. These critiques, together with a good deal of recent evidence, have sustained the theoretical and empirical resilience of the concept of response styles.

Although there are potentially contentious points at the theoretical level, most of the literature concerned with the content-style issue admits to the possible existence of response styles but disagrees on the interpretation of the evidence for their importance. The basic controversy is thus at the level of data. And the data most highly disputed in the response style literature are those of acquiescence (Messick, 1967).

Consequently, we now turn to the question of the existence of acquiescence. Prior to proceeding to the empirical level, some distinctions are proposed that have been made before (Jackson & Messick, 1965) and that appear to be useful for other investigators (Campbell et al., 1967).

NATURE OF ACQUIESCENCE

Example

Suppose each of several subjects is asked to indicate what kind of a person he is by responding to phrases. If he interprets the phrase as describing him accurately, he should respond "true"; if he does not consider the phrase as an accurate self-description, he should respond "false." The subjects might be presented with four phrases, consisting of two adjectives and their negations: "happy," "sad," "not happy," and "not sad." Thus the subjects would respond to four items, as indicated in Table 1. For each of the items, each subject responds "true" or "false," and corresponding to his responses, let us give him a score of 1 for every phrase endorsed, or marked "true," and a score of 0 for every phrase not endorsed, or marked "false."

These four items were chosen to represent a content dimension corresponding to a positive versus negative feeling state. It is now possible to compute the six different correlations between pairs of the four items, as listed in Table 2. If the subjects are responding consistently to the *content* of the items, one would expect the signs of the respective correlations to be as presented in the Content column of Table 2. Thus, for example, we would expect response to "happy" and "sad" to correlate negatively, since these adjectives are usually conceptualized to be at opposite poles of the same content dimension. In contrast, the pair "happy"–"not sad" should correlate positively, since persons who thought of themselves as "happy" would presumably also consider themselves "not sad," while some persons would respond "false" to both phrases. A rather different set of item responses, with its own consequent pattern of correlations, would be observed if the subjects are responding to *noncontent* attributes of the items. It is the acquiescence response style that concerns us here.

TABLE 1

SUBJECT RESPONSES TO FOUR ITEMS

Item number	Item	Key	
1	Happy	T = 1	F = 0
2	Sad	T = 1	F = 0
3	Not happy	T = 1	F = 0
4	Not sad	T = 1	F = 0

Note.—Abbreviations are: T = true; F = false.

TABLE 2

EFFECTS OF CONTENT AND ACQUIESCENCE

Item number	Correlation pair	Content	Acquiescence	
			Agreement	Acceptance
1, 2	Happy–Sad	−	+	+
1, 3	Happy–Not happy	−	+	−
1, 4	Happy–Not sad	+	+	−
2, 3	Sad–Not happy	+	+	−
2, 4	Sad–Not sad	−	+	−
3, 4	Not happy–Not sad	−	+	+

Two Types of Acquiescence

The first type of noncontent response is that of *agreement acquiescence*. This response style can be defined as individual differences in the tendency to agree. In the personality domain, the definition would include agreement with personality statements. Thus a person responding solely in terms of agreement acquiescence would respond "true" to all four items; a person at the negative pole of this dimension would respond "false" to all four items. If only agreement acquiescence were determining responses to the items, then it is clear that all items would be positively correlated—even the ones opposite in content. This correlational expectation is presented in the Agreement column of Table 2. It can be seen that the pattern of correlations to be expected from content responding is radically different from that expected from agreement acquiescence alone. The word "agreement" is used to be consistent with Couch and Keniston (1960) and with Messick (1967).

The second type of noncontent response is that of *acceptance acquiescence*. This response style can be defined as individual differences in the tendency to consider characteristics as descriptive. In the personality domain, this definition would include accepting characteristics as self-descriptive. Thus an extreme accepter would say "true" to all items that describe personality characteristics and would say "false" to all items that deny such characteristics; an extreme rejecter would respond in the opposite fashion, disclaiming any and all characteristics. Thus an accepter would say "true" to "happy" and "sad," but "false" to "not happy" and "not sad." The rejecter would say the reverse. If only acceptance acquiescence were determining responses to the items, the pattern of correlations among the items would be quite different from those previously presented. The only positive correlations would be among the two pairs that represent consistent presence or absence of a trait, such as "happy"–"sad" and "not happy"–"not sad," since persons would respond either "true" or "false" to both. Positively and negatively worded pairs would correlate negatively as shown in the final column of Table 2. The concept of acceptance acquiescence was first delineated in 1965 (Jackson & Messick, 1965).

Background for the Distinction

Acquiescence was initially defined as a tendency to respond "true" to test items (e.g., Cronbach, 1942). Jackson and Messick reasoned that if acquiescence were simply a process of "true" responding (agreement), and if acquiescence accounted for any significant portion of variance on test scores, this agreeing tendency should manifest itself as a factor separating the true-keyed and false-keyed subscales of various tests in a factor analysis of such devices. Jackson and Messick (1961, 1962a, 1962b) obtained almost perfect verification of this notion with the MMPI in three different samples. Block (1965), however, argued that the existence of this factor did not rule out a content interpretation. Finally, item-reversal studies were carried out (Lichtenstein & Bryan, 1965; Rorer & Goldberg, 1965a, 1965b). Items were written that were the reverse of the originals (as "sad" or "not happy" may be considered content opposites of "happy," in our example), and the correlations between original and reversed items were examined. The investigators reported correlations between original and reversed items and scales almost as high as estimates of item and scale reliabilities, a finding interpreted as apparently indicating strong content consistency, although content was not otherwise defined. Data such as these, as well as data showing in certain cases low intercorrelations among measures of acquiescence taken from different tests and lack of consistent correlates of acquiescence with external criteria, led Rorer (1965) to the conclusion that acquiescence was a myth. This conclusion unfortunately left certain critical features of previously published data unexplained, such as the perfect factorial separation of true- and false-keyed MMPI scales.

Problem with reversal studies. Two distinct types of item reversals were used in the reversal studies cited above. One may be called a polar-opposite form of content reversal, the other a negation form. In the example cited in Tables 1 and 2, the pair "happy"–"sad" would represent polar opposites on a content dimension, and thus "sad" could be considered a polar-opposite reversal of "happy." In contrast, "not happy" would represent a negation form of content reversal, since the phrase does not define the opposite pole of the content dimension. With items scored as indicated in Table 1, it is quite clear that the correlation for both types of reversal would be negative in sign if content were determining subjects' responses. This effect was seen, for example, in the first two rows of Table 2. It is also obvious, as shown in Table 2, that the expected correlation from the agreement hypothesis is opposite in sign. Since reversal studies reported a reasonably high negative correlation for both types of item reversal, it was clear that agreement acquiescence must be a relatively minor determiner of responses on inventories such as the MMPI.

However, there remained a major problem. The high negative correlations observed in reversal studies would be equally consistent with an acceptance hypothesis as with a content hypothesis if negation reversals were used in these studies; this effect is illustrated in row 2 of Table 2, for example. An examination of the various reversal studies made it obvious that a large proportion of their reversed items were negation reversals, rather than polar-opposite reversals. For example, over 100 of Rorer's were reversed by prefixing the phrase "It is not true that . . . ." Thus the existence of strong "content" correlations could provide equivalently good verification for acceptance as for content. Studies which seemed perfect to many thus appeared to have a flaw. These issues were reported quite some time ago (Jackson & Messick, 1965), but to date only a few others have taken up our point of view (Campbell et al., 1967). Most investigators continue equating all acquiescence with agreement and uncritically interpret its complement as "content" (e.g., Bock, Dicken, & Van Pelt, 1969).

Ubiquitous MMPI. A number of questions about the MMPI remained unanswered by content proponents. Factor analyses of the MMPI have typically found two major common factors and a number of minor factors. Content proponents have assumed these factors to reflect different aspects of content (e.g., Block, 1965; Wiggins, 1966). Thus Rorer (1965), for example, concluded that "it is an arbitrary decision to label the factors in terms of response sets, biases, or styles, rather than in terms of the criterion groups which the MMPI scales were originally devised to discriminate [p. 142]." But what could be the nature of this content, which *despite* the different content names given it (e.g., depression, schizophrenia), appeared to implicate virtually all MMPI items and scales in only two major common dimensions? Clearly the two major MMPI factors represent such a general kind of "content" as to be something quite different from the content assumed for it by the constructors of the MMPI. The definition of content as implied in the use of different construct or trait names for the scales would seem to require some convergent and discriminant properties, a property in the MMPI notable for its rarity. One of the content factors would require every true-keyed subscale to measure something that is different from and opposite to what every false-keyed subscale measures. Furthermore, the same trait would have to be measured by all true- and false-keyed subscales. A content interpretation would require an explanation for this single dimension in which the true-keyed scales when keyed positively for content measured one pole, and the false-keyed scales when similarly keyed positively for content measured the opposite pole. This would be analogous to a true–false ability test in which all items keyed "true" for a correct response and all items keyed "false" for a correct response load the same factor in opposite directions. A hypothesized general ability factor would ordinarily have all correct-keyed items loading in the same direction!

Subportions of MMPI scales, such as those identified by true- and false-keyed items, would seem of necessity to correlate highly in the content direction; yet this was frequently not the case. Indeed, it was clear that subscales added together to form a total score would possess components showing high negative relationships, a fact also revealed by the negative relationships between a number of true- and false-keyed MMPI scales (Jackson & Messick, 1961). Consequently, the different content scales did not provide their own factors in factor analysis.

Another MMPI factor also implicates every scale, seemingly in terms of an important connotative stylistic attribute of responses to personality items, namely, their desirability: The correlation of an MMPI scale with this factor is easily predictable from knowledge of the social desirability scale values of its items. Yet this relation must also be explained by content. In addition, MMPI scales are organized in an interesting circular pattern in this two-factor space, with the loading on the factor separating true- and false-keyed scales systematically decreasing as scales become more extreme in desirability. This organization of subscales within a factor structure dominated by two large factors is not consistent with any published account of the rationale on which scale construction for the original MMPI was based, nor have we encountered any theorist who could provide a plausible content interpretation for this circular pattern. Block's (1965) reconsideration of the MMPI content as representing very broad dispositions did not appear to explain this data on the internal structure of the MMPI. It was not a relevant rejoinder to this stylistic interpretation of the MMPI internal structure to demonstrate that there are possible external correlates of these two large dimensions, because, as noted earlier, such a finding would be just as consistent with a stylistic interpretation (Jackson & Messick, 1958).

The conventional definition of content (Campbell & Fiske, 1959) requires that distinct measures of separate traits be not so highly correlated as to preclude their statistical and substantive independence. Personality scales failing to show appropriate discrimination in terms of internal structure can hardly be expected to be associated with independent constructs. Neither can they be dignified as representing independent content nor legitimately be referred to by different names.

Factor analyses of Rorer's reversed form of the MMPI (Jackson & Messick, 1965), performed separately for males and females, again verified the existence of the two major factors, including the true–false factor. Even though the scales were reversed in direction, the same factorial separation of true- and false-keyed scales existed. The loadings of true-keyed scales on one of the two major dimensions were opposite in sign to those of the false-keyed scales. The usual type of two-factor pattern, taking account of appropriate keying reversals, resulted when a Tucker (1958) interbattery factor analysis was undertaken of the original and reversed forms together. This analysis, which focuses only upon the variance in common to original and reversed forms, was undertaken to completely avoid the problem of item overlap. The results of the analyses were essentially identical to those previously reported; indeed, the first two factors were more massive than ever, accounting for 94% of the total variance.

Taken together with the previously presented information that the reversal data were quite amenable to a style interpretation, this interbattery analysis is also consistent with the style position. The existence of only two major factors to account for the correlations between original and reversed scales—to account for virtually all of the observed response consistency between the two forms—would seem to violate reasonable requirements for a content position that attributes distinct content to different scales.

In the remainder of the review we turn to fresh data to evaluate the two acquiescence styles described in this section. It is appropriate to present a few highly pertinent studies in some detail, in order to avoid a superficial treatment of masses of data.

RECENT EVIDENCE FOR ACQUIESCENCE

Existence of Noncontent Responding

The types of endorsement patterns to be expected for phrases and their polar opposites or reversals were described previously, and the effects of content and the two types of hypothesized acquiescence were summarized in Table 2. The types of correlations to be expected from a content position as well as from the two acquiescence positions were described there for each of the various possible kinds of item comparisons. The most definitive contrast to be expected from a content and a noncontent position can be seen to be in the comparison of polar-opposite terms. A content position would indicate that polar opposites should correlate highly negatively, while both the agreement- and acceptance-acquiescence positions would demand positive correlations among polar opposites. Thus a study dealing with polar opposites presented in the fashion of the items of Table 1—for instance, adjectives—would yield quite contrasting results from the content and style positions. Such a study could not disentangle the two types of acquiescence, but it could demonstrate the existence of acquiescence per se.

The correlational expectations summarized in Table 2 hold only for the ideal case of either content or acquiescence responding. If both content and acquiescence responding influenced the results as components of variance, the expected correlation between polar opposites would be some compromise between negative correlation associated with content and positive correlation resulting from acquiescence. The correlation might even be zero! A zero correlation by itself would be difficult to interpret, however, since it could be due either to error of measurement or to some compromise of the two response dispositions. As a result, it was necessary to design a somewhat more complicated study than the simple and obvious one.

The most agreed-upon expectations regarding the content correlation of a set of scales could be derived from the widely cited measurement work of Osgood, Suci, and Tannenbaum (1957). These authors established that evaluation, activity, and potency define major connotative dimensions in an individual's description of common objects. Bentler (1969) hypothesized that acquiescence tendencies could be described clearly in the domain of connotative meaning. He gathered adjectives of the polar-opposite type (e.g., "good," "bad") to assess each pole of the three semantic differential factors. Thus, six sets of adjectives were collected: positive evaluation (E +), negative evaluation (E −), positive activity (A +), negative activity (A −), positive potency (P +), and negative potency (P −). An adjective checklist form of the semantic differential was constructed, on which a given person's concept score for each pole of each dimension was the number of adjectives checked. One hundred concepts were rated by 200 undergraduates. From these data, a scale of 25 adjectives was selected by monotonicity analysis (Bentler, 1970) to assess each of the six poles mentioned above. The number of adjectives (out of 25) checked by a person for a concept on one of the six scales represented the concept score for a given person. These total scores formed a 200 × 6, Person × Scale, data matrix. The six scales were correlated, yielding a set of relationships among each of the scales.

What type of results would one expect if only the specific semantic content determined a scale's total score—that is, what should the correlation matrix look like if paper-and-pencil tests are unaffected by any type of acquiescence? Obviously, if two scales for a given semantic factor measure opposite ends of the same factor, these opposites should be highly negatively correlated. Thus, E + and E − should be highly negatively related, as should A + and A −, and P + and P −. Cross-factor correlations should be reasonably low—certainly lower in absolute value than the absolute value of the correlations of polar opposites. Alternatively, if either agreement or acceptance acquiescence is a major determiner of scale scores, polar opposites would tend to be less than perfectly correlated since stylistic and content response tendencies would be working in opposition. Acquiescence as a stylistic characteristic would be assessed by the total number of adjectives checked across all six scales. It is not possible to separate the two types of acquiescence here, so that this score represents the operation of both processes in unknown combination. Yet the well-established bipolarity and relative orthogonality of the factors on which the scales were based would largely exclude the possibility of a subject obtaining an extremely high acquiescence score on the basis of content considerations alone. If only acquiescence determined responses, all scales, even including polar opposites, would be significantly positively related. If specific content also affected responses, the correlation of polar opposites would represent some compromise between these two opposing correlational tendencies.

The correlations among the six scales in the adjective semantic differential study are presented in Table 3. The correlations among polar opposites are in all three cases close to zero, but positive in sign. Furthermore, all 12 correlations between subscales from one content area and subscales from another content area are larger in absolute magnitude than the three correlations representing polar opposites of the same scale. Thus, the correlation between two subscales, A + and E +, representing different content, is relatively high (.65), while the correlation between two scales designed to measure the same content, E + and E −, is low and positive (+ .03), a direction opposite to that expected from the hypothesis of consistent responses to content. Thus, these results provide strong support for an acquiescence hypothesis and are not consistent with the hypothesis that content is the sole or even the major determinant of responses to descriptive adjectives.

Even better evidence favoring an acquiescence hypothesis can be obtained from the interrelations among the scales. If acquiescence, as well as content, accounts for the pattern of correlations in Table 3, then partialing out the effects of acquiescence from the correlations should yield a pattern of relationships which corresponds to expectations based on content-determined responding. Using as a measure of acquiescence the total number of adjectives checked across all six scales (which, incidentally, yielded a coefficient alpha internal

TABLE 3

Semantic Adjective Differential Correlations

Adjective set	E+	E−	A+	A−	P+	P−
E+	1.00					
E−	.03	1.00				
A+	.65	.22	1.00			
A−	.23	.63	.08	1.00		
P+	.30	.57	.59	.35	1.00	
P−	.50	.39	.32	.64	.18	1.00

TABLE 4

Semantic Adjective Differential Correlations, Acquiescence Partialed Out

Adjective set	E+	E−	A+	A−	P+	P−
E+	1.00					
E−	−.76	1.00				
A+	.33	−.50	1.00			
A−	−.40	.33	−.77	1.00		
P+	−.36	.15	.15	−.28	1.00	
P−	.06	−.16	−.38	.32	−.70	1.00

consistency of .79), partial correlations were computed among the scales holding constant the effects of acquiescence. These results are presented in Table 4. In contrast to the results presented in Table 3, the pattern of correlations in Table 4 corresponds to what one would expect on the basis of content responding—polar opposites correlate highly negatively, and the absolute values of all three polar-opposite correlations are larger in magnitude than the 12 cross-scale correlations. A more complete verification of the hypothesized acquiescence tendency could hardly have been expected.

An additional study[3] verified the applicability of the acquiescence process demonstrated for semantic meaning to the domain of personality assessment. A total of 193 mothers of children in the age range 0–6 years were asked to describe their children using the adjectives of the adjective semantic differential. Total scores for each of the two poles of the three semantic factors were obtained and intercorrelated, in a fashion quite similar to that described in the previous study. The lower left-hand portion of Table 5 presents the correlations among the various subscales. Once again, one finds correlations very close to zero between polar opposites, as well as fairly consistent positive correlations among subscales from different content domains. A total adjective acquiescence score was obtained by counting up the total number of adjectives checked by a mother in describing her child. The intercorrelations among the various subscales, which result when the adjective acquiescence scores are partialed out, are presented in the upper right-hand portion of Table 5. It can be seen that the partial correlations once again correspond much more accurately in size and sign to what would be expected on the basis of content responding.

A final study bearing on the presence of acquiescence emerged as a by-product of an investigation which sought to evaluate empirically two alternative methods in multivariate attitude scaling. Boyd and Jackson (1969) first applied the technique of multidimensional successive intervals scaling to a set of 18 attitude items hypothesized to reflect bipolar dimensions of Authoritarian Rejection, Power and Toughness, and Superstition. In addition, they employed a procedure in which behavior descriptions of imaginary persons designed to be high and low on each of these dimensions were incorporated into the scaling procedure. Multidimensional scaling resulted in the finding that both items and hypothetical persons were consistently represented as predicted on each of the three dimensions, without any

TABLE 5

Child Adjective Description Correlations and Partial Correlations

Adjective set	E+	E−	A+	A−	P+	P−
E+		−.60	−.04	−.10	−.46	.02
E−	−.18		−.29	.30	.17	−.07
A+	.35	.13		−.53	.10	−.45
A−	.15	.44	−.09		−.29	.13
P+	.14	.44	.54	.12		−.67
P−	.25	.15	−.01	.28	−.10	

Note.—Correlation are in the lower left, partial correlations are in the upper right.

---

[3] Bentler, P. M. Children's personalities assessed by an adjective check list. Unpublished manuscript, University of California, Los Angeles, 1969.

TABLE 6

Description of Scales Used in Analyses

Content scales

Label	Content	Format	Wording	Keying	KR-20[a]
ESPT	Exhibition	Self-descriptive	Positive	True	.23
ESPF	Exhibition	Self-descriptive	Positive	False	.15
ESNT	Exhibition	Self-descriptive	Negative	True	.33
ESNF	Exhibition	Self-descriptive	Negative	False	.46
EAPT	Exhibition	Attitude	Positive	True	.07
EAPF	Exhibition	Attitude	Positive	False	.39
EANT	Exhibition	Attitude	Negative	True	.11
EANF	Exhibition	Attitude	Negative	False	−.05
PSPT	Play	Self-descriptive	Positive	True	.38
PSPF	Play	Self-descriptive	Positive	False	.42
PSNT	Play	Self-descriptive	Negative	True	.22
PSNF	Play	Self-descriptive	Negative	False	.25
PAPT	Play	Attitude	Positive	True	.34
PAPF	Play	Attitude	Positive	False	.35
PANT	Play	Attitude	Negative	True	.10
PANF	Play	Attitude	Negative	False	.33
SSPT	Succorance	Self-descriptive	Positive	True	.31
SSPF	Succorance	Self-descriptive	Positive	False	.32
SSNT	Succorance	Self-descriptive	Negative	True	.26
SSNF	Succorance	Self-descriptive	Negative	False	.51
SAPT	Succorance	Attitude	Positive	True	.08
SAPF	Succorance	Attitude	Positive	False	.32
SANT	Succorance	Attitude	Negative	True	.13
SANF	Succorance	Attitude	Negative	False	.19
USPT	Understanding	Self-descriptive	Positive	True	.38
USPF	Understanding	Self-descriptive	Positive	False	.35
USNT	Understanding	Self-descriptive	Negative	True	.28
USNF	Understanding	Self-descriptive	Negative	False	.46
UAPT	Understanding	Attitude	Positive	True	.42
UAPF	Understanding	Attitude	Positive	False	.38
UANT	Understanding	Attitude	Negative	True	.09
UANF	Understanding	Attitude	Negative	False	.36

Response-style marker scale

Label	Content	Source of items	Wording	Keying	KR-20[a]
HMPT	Heterogeneous	MMPI	Positive	True	.51
HMPF	Heterogeneous	MMPI	Positive	False	.65
HMNT	Heterogeneous	MMPI	Negative	True	.45
HMNF	Heterogeneous	MMPI	Negative	False	.52
HPPT	Heterogeneous	PRF	Positive	True	.44
HPPF	Heterogeneous	PRF	Positive	False	.33
HPNT	Heterogeneous	PRF	Negative	True	.31
HPNF	Heterogeneous	PRF	Negative	False	.25
FRT		F Scale	Relative	True	.46
FRF		F Scale	Relative	False	.25
FAT		F Scale	Absolute	True	.50
FAF		F Scale	Absolute	False	.17
DA		Desirability Scale (PRF Form A)		Balanced	.55
DB		Desirability Scale (PRF Form B)		Balanced	.61

TABLE 6—(*Continued*)

Adjective checklist and other variables		
Label	Description	KR-20[a]
AD	Desirable adjectives	
AN	Neutral adjectives	
AU	Undesirable adjectives	.12
INFR	Infrequency Scale	
SEX	Sex of subject	

[a] KR-20 refers to Kuder Richardson Formula 20 internal consistency reliability.

exceptions and with no reversals. This indicated that judges agreed that the items substantively defined appropriate poles of each of the three dimensions. However, when these same items were administered in a response format and factor analyzed, results were equivocal, primarily because of the intrusion of massive levels of acquiescence. The positive and negative items were all scored positively to obtain a total-agree score, essentially canceling the effects of content. When biserial correlations were obtained between each item and the total number of agree responses, 17 of the 18 biserial correlations were positive, 13 significant at the .05 level. If responses were made solely on the basis of content, all of these correlations should have hovered around zero. Acquiescence bias in a response format can thus be elicited in a set of attitude items that have been shown in a judgmental scaling study to be associated uniquely with a set of clear, bipolar content dimensions; its presence produces a consequent attenuation in the reliability and factor clarity of responses.

Existence of Two Processes

The studies cited above provide clear-cut evidence for the operation of acquiescence, but they provide no data relevant to the separability of the two types of acquiescence. A dissertation completed by Morf,[4] and described by Morf and Jackson (in press), shows quite clearly the operation of each of the acquiescence processes. Morf and Jackson used a Guttman (1958) facet design with items carefully selected to elicit hypothesized content and stylistic processes, specially constructed response-style marker scales, factor analysis, and an objective rotation procedure that would yield an internal analysis of the structure of the items which would avoid many of the problems inherent in the use of the MMPI. This type of design allows content to appear under a variety of types of item formats, but also allows response styles to appear in the context of a variety of content dimensions if response styles exist. None of these properties can safely be assumed to be independent or systematically varying on the MMPI, where even the identification of content poses some problem.

A two-part questionnaire consisting of 560 true–false personality items and an adjective checklist were administered to 87 male and 108 female university student volunteers in two sessions. These items were selected according to very exacting criteria so as to provide measures of eight possible response processes that might be operating: agreement acquiescence, acceptance acquiescence, desirability, adjective acquiescence, and four content dimensions—Exhibition, Play, Succorance, and Understanding. It was decided to treat adjective acquiescence as potentially distinct since it is clear from the analysis presented above that adjective acquiescence could represent either agreement or acceptance acquiescence, or some blend of the two. In addition, adjective and statement formats were considered sufficiently different to allow for the appearance of a distinct method factor.

Items were created for each of the four content dimensions according to a very restricted design. First, items were of two broad types: personality statements and attitude statements. Personality statements were designed as self-descriptive first-person statements, while attitude statements were written to elicit an evaluative response toward an external referent. Second, for each of the two types of items, the statements were worded positively or negatively, in accord with the acceptance acquiescence hypothesis described previously. Third, for each type of wording, items were written so as to be true- or false-keyed in the content direction. Thus each content dimension was represented by eight scales. Each of the eight scales was itself composed of from five to eight items. In addition to the item attributes described above, each of the items was selected so as to have intermediate endorsement frequencies, low correlations with desirability, and moderate content saturations as reflected in biserial correlations with scale score totals which had been obtained in other studies. The 32 content variables which resulted from this design are listed in the first section of Table 6, together with their internal consistencies.

---

[4] Morf, M. E. An analysis of two response sets: True responding and item endorsement. Unpublished doctoral dissertation, University of Western Ontario, 1969.

A number of response-style marker scales were also included in the analysis. Four heterogeneous 30-item acquiescence scales were constructed with MMPI items neutral in desirability and intermediate in endorsement frequency. As before, these scales were worded positively or negatively and keyed true or false. A similar set of four Personality Research Form (PRF) acquiescence scales was constructed from the original, prepublication pool of items of the PRF (Jackson, 1967a). The Clayton and Jackson (1961) absolutely worded and relatively worded true- and false-keyed attitude scales based on the California F Scale also generated four acquiescence scales. Finally, the Desirability scales from the PRF were also included in the analysis. The 14 response-style marker variables just described are listed in the second section of Table 6, which also lists three adjective checklist variables. The adjective checklist consisted of 45 desirable, 90 neutral, and 45 undesirable adjectives selected from Anderson's (1968) list; they were arranged in random order. Separate scores were obtained to yield the number of desirable, neutral, and undesirable adjectives checked as self-descriptive. Finally, as shown in Table 6, there were two irrelevant variables, an Infrequency scale and the sex of the subject.

Thus, there were 51 variables in all. The intercorrelations among the 51 variables were computed; eight principal axes factors were extracted from the correlation matrix to be in agreement with the number of response processes hypothesized to be operating, and the initial factor matrix was rotated in a preliminary analysis by Varimax, and for the final reported solution by Clustran (Bentler, 1971) to the factor position hypothesized at the start of the study. The eight factors were correlated, so Bentler's (1968) factor contribution matrix was computed. This matrix represents the orthogonal factor solution that is closest in the least-squares sense to the oblique solution and permits assessing the percentage of variance attributable to each factor.

Before moving on to a description of the factors, a procedural word seems in order. Human intervention in the analysis process is kept at a minimum by the procedures utilized. Once items were constructed for the 51 variables, no additional subjective decisions remained to bias the outcome beyond the "bias" introduced in hypothesizing eight response processes. The subjects determined the correlation matrix; the number of factors was ordained, the hypothesized position of variables in the space automatically determined the multidimensional position of the final factors, the correlation among the oblique factors was also automatically determined without intervention. As a check on the possibility that Morf and Jackson's (in press) reporting of results based on the Clustran rotation, which involved hypothesizing a particular structure, might have introduced a source of bias, these results were compared with those based on a normalized Varimax rotation, and with the results as reported by Morf (see Footnote 4); the original data contained minor inaccuracies in item keying. There was a very high degree of agreement between these solutions. The interpretation of the eight primary factors was identical.

The orthogonal factor-loading matrix representing the factor contributions is presented in Table 7. Like any factor-loading matrix,

TABLE 7

Rotated Factor Matrix

Variable	Factor								
	I	II	III	IV	V	VI	VII	VIII	h^2
ESPT	23	23	−16	−06	55	10	04	−10	46
ESPF	−13	−17	15	−21	54	−02	11	−15	44
ESNT	12	−08	13	−04	57	08	14	−11	40
ESNF	−25	21	08	06	47	28	09	09	43
EAPT	19	10	−18	−11	41	12	−10	−09	29
EAPF	−33	−11	04	05	40	05	26	07	36
EANT	34	−08	−06	−01	25	−06	00	−01	19
EANF	−25	09	08	10	31	07	−01	11	20
PSPT	22	14	14	24	15	41	13	03	35
PSPF	−25	−09	−02	−06	12	60	08	−20	50
PSNT	33	−01	−05	13	09	63	01	−08	54
PSNF	−26	26	−11	02	17	42	07	00	36
PAPT	21	22	07	02	05	47	13	−08	34
PAPF	−33	−15	−16	02	02	43	−06	−09	35
PANT	29	−11	−12	10	00	57	−03	−09	45
PANF	−33	12	−01	12	12	45	13	02	37
SSPT	04	16	−27	11	08	15	55	−17	47
SSPF	−31	−09	06	06	14	−07	52	01	41
SSNT	11	−04	05	06	17	04	55	−07	36
SSNF	−08	26	−22	00	04	18	51	06	42
SAPT	31	17	−12	13	09	04	39	−13	33
SAPF	−40	−15	18	02	19	−13	48	09	51
SANT	15	−11	10	−02	−06	15	44	07	27
SANF	−20	04	−08	−04	−05	09	49	−08	31
USPT	12	31	02	09	11	−09	−07	54	44
USPF	−22	05	−06	−15	01	04	−02	68	54
USNT	12	−22	00	09	01	−18	−09	45	31
USNF	−33	20	10	−06	06	05	−19	46	42
UAPT	30	13	13	−07	−14	−25	−07	39	37
UAPF	−30	−12	−08	−11	−09	−01	15	66	59
UANT	20	−18	−05	−06	−11	07	−04	17	13
UANF	−40	11	18	−22	−04	−10	00	46	48
HMPT	34	53	−24	20	13	00	08	05	52
HMPF	−30	−52	49	−19	−01	−04	−05	−09	65
HMNT	33	−36	41	−07	06	−08	−22	00	47
HMNF	−31	50	−24	05	08	10	23	−06	48
HPPT	43	40	22	12	12	−05	−04	09	43
EPPF	−33	−24	−03	−08	−12	−14	−10	−28	30
EPNT	28	−25	−30	−11	−07	06	−13	−32	37
HPNF	−31	38	22	04	00	16	02	10	33
FRT	42	27	05	−18	−22	05	08	−38	49
FRF	−37	10	−09	−14	−03	12	06	−21	24
FAT	49	26	10	−23	−22	00	09	−35	55
FAF	−39	17	−04	−28	−06	03	16	−10	30
DA	−10	−10	65	08	05	−11	−05	11	48
DB	−02	−11	68	−15	−01	−05	09	02	51
AD	05	11	56	43	01	08	−09	−07	53
AN	10	26	−13	68	−11	11	11	−05	60
AU	05	21	−51	47	−01	06	13	−09	56
INFR	−16	−09	−12	10	18	04	−14	−01	11
SEX	−05	06	−13	−03	−24	09	34	13	22
Sum of squares	3.74	2.48	2.70	1.54	2.15	2.52	2.58	2.80	
Percentage of common variance	18.2	12.1	13.2	7.5	10.5	12.3	12.6	13.6	

Note.—Decimals omitted. This matrix is the Clustran matrix providing the least squares orthogonal fit to the hypothesis factors.

its entries represent the correlation of variables with factors. We deal first with the stylistic factors, later with the content factors.

1. The first factor is clearly an agreement acquiescence factor. All true-keyed scales loaded positively on the factor, and all false-keyed scales loaded negatively. This separation of true- and false-keyed scales occurred regardless of the positivity or negativity of wording. It might be noted that the highest loading scales are attitude scales, rather than personality scales, indicating the apparently greater appearance of agreement acquiescence in the attitude domain, rather than the personality domain.

2. Positive loadings on the second factor are of two types: those that reflect the endorsement of a positively worded item and those that reflect denial of a negatively worded item. These are equivalent, and they represent the unequivocal operation of acceptance acquiescence. It is particularly noteworthy that of the 44 personality and attitude scales in which direction of wording and direction of keying are systematically varied, all but one loaded in the appropriate direction. The one exception to perfect loading pattern was a negligible .05 loading for the positively worded false-keyed Understanding self-descriptive content scale. The strength of the pattern effect can be seen in the second factor column of Table 7 by noting the almost perfect repetitive pattern of signs in the sequence (+ − − +). A biserial correlation coefficient computed between the magnitude of the factor loading and the expected direction of loading was also computed; its value was .99.

3. This factor is a rather clear-cut desirability factor. The highest positive loadings were obtained for the parallel PRF Desirability scales (.68, .65), and the highest negative loading was obtained for the number of undesirable adjectives endorsed (− .51). Thus the existence of this readily identifiable factor casts doubt on Block's (1965) contention that desirability responding can be explained in terms of consistent responding to pathological content, since the variables just mentioned were composed of items from nonpathological content domains.

4. The neutral, undesirable, and desirable adjective scales obtained the highest loadings on this factor (.68, .47, .43), with the highest loading being obtained by neutral adjectives. Thus, this factor is clearly identifiable as adjective acquiescence.

5–8. The next four factors represent the content factors Exhibition, Play, Succorance, and Understanding. The high loadings obtained for the various content variables designed to assess each of the factors can easily be seen in the appropriate column of Table 7. The first letter of the variable's identification code indicates its content; thus the E variables load highly on Factor V, P variables on VI, and so on. It should be pointed out that these content factors are clearly identifiable even though only relatively short scales comprised of items with moderate content saturation were used—too low to permit their inclusion in the final PRF—suggesting that when content is truly identifiable as such, even in the context of an investigation designed to identify and label response styles, it will define its own factors in spite of the operation of the response styles.

Thus it has been possible to identify response styles and content unambiguously using objective multivariate procedures. Although it is not necessary to describe this study further to meet the objectives of this section of the review, namely, the empirical identification of two distinct acquiescence processes, an understanding of the processes themselves will be enhanced by looking more closely at the correlations among the oblique factors that were obtained.

The 8 × 8 correlation matrix of the oblique factors was resolved into uncorrelated principal components, and four components were rotated using a Varimax procedure. The second-order loading matrix that resulted from this procedure is presented in Table 8. A number of interesting features can be seen in that table. Acceptance acquiescence defines its own factor (I), but this factor also has a high loading from adjective acquiescence, indicating that adjective acquiescence appears to represent an acceptance process rather than an agreement or true-saying process. It will be remembered that the adjective presentation format, by itself, did not allow one to decide which type of acquiescence might be operating.

Agreement acquiescence also defines a factor (III), has virtually no loading from other style factors, and, interestingly, has a high negative loading from the Understanding content factor. Thus agreement acquiescence would seem to be associated negatively with a dimension of intellectual curiosity. Desirability also defines its own factor (IV). Finally, the content scales appear to merge at the second-order level into one general content factor (II), with the three interpersonal dimensions showing high positive loadings, and Understanding loading negatively.

Thus, the second-order analysis provides support for the concept of two relatively independent acquiescence processes: agreement and acceptance. This conclusion is consistent with the low correlation (.11) that had been obtained between the corresponding primary factors.

TABLE 8

Rotated Second-Order Factor Matrix

First-order factor	Second-order factors			
	I	II	III	IV
I Agreement acquiescence	24	−26	**76**	−05
II Acceptance acquiescence	**90**	12	−10	14
III Desirability	−16	−02	−09	**−92**
IV Adjective acquiescence	**84**	13	27	06
V Exhibition	−05	**77**	−02	−19
VI Play	29	**69**	15	00
VII Succorance	04	**68**	−09	36
VIII Understanding	10	−34	**−80**	−15

Note.—Decimals are omitted. Salient loadings are in boldface.

Masking of Content

It is of some importance to reiterate that response styles can mask the true existence of content. The studies described in the previous two sections have already attested to the masking effect. The expected content correlations among semantic differential scales only appeared after an appropriate correction for the operation of acquiescence. While no content interpretation of this correction seems at all plausible to us, it should be pointed out that even if it were possible to make a post hoc content interpretation of this correction, no content hypothesis yet published would have led an investigator to worry about the correction in the first place. In a content area less clear-cut than that of the semantic differential, an investigator not concerned with acquiescence would accept his "content" correlations at face value—yet possibly be tremendously misled. For example, virtually all constructors of adjective checklists do not bother to worry about acquiescence—remember, they have even been advised not to worry about it. Consequently, one finds such surprising results as content correlations among presumably different traits on the order of .92, .86, and .81; these were the correlations reported among measures of anxiety, depression, and hostility in a study with the Multiple Affect Adjective Checklist (Lubin & Zuckerman, 1967). As we attempted to detail before, we can see no legitimate reason to dignify "content" traits with names implying independent constructs when the trait correlations are as high as this, and *particularly when an acquiescence hypothesis would lead one to expect distortion in a direction of high positive correlations*.

The Morf and Jackson (in press) study described in the previous section allowed for the identification of content, despite the operation of response styles; clear-cut results like this have only rarely been published (e.g., Messick, 1962). But what if content and style had been perfectly confounded, making it impossible to decide whether an observed factor was that of content or style? And for that matter, what if content and style were so confounded that the content factors became so highly correlated that they merged one into the other? Clearly, one can conceive of conditions where response styles make the identification of content difficult because of the greater preponderance of stylistic variance. For example, in a study by Trott and Jackson (1967), the emergence of an acquiescence factor was found to be strongly associated with the speed of presentation of personality items. When items were presented under relatively speeded conditions, the largest factor obtained indicated an almost complete separation of true- and false-keyed scales, while under relatively unspeeded conditions, the acquiescence factor was only the sixth largest and was not as clearly defined. Similarly, content factors were decidedly more well-determined in the less speeded condition.

A further study bearing on the importance of content saturation in relation to acquiescence in personality items was reported by Jackson and Lay (1968). Jackson and Lay identified three factors as being primarily attributable to response style. One was predominantly associated with the familiar tendency to endorse desirable or undesirable scale content, while a second factor was associated with a tendency to endorse many or few adjectives, identifiable in the present context as adjective acquiescence. A third response-style factor was also clearly identifiable as acquiescence: for this factor, positive loadings emerged for scales composed of heterogeneous MMPI and California Psychological Inventory (CPI) items all keyed true, and negative loadings for MMPI and CPI items keyed false. However, items high in content saturation derived from the PRF (Jackson, 1967a) showed virtually no tendency to load any of the style factors nor was there evidence for any style influence on these content-saturated items. This finding is consistent with the evidence from Trott and Jackson (1967), which indicates that a trading relation exists between acquiescence to personality items and content saturation. When content saturation is high, acquiescence is low, and vice versa. One conclusion relevant to scale construction that can be drawn from this study is that an efficient means for reducing acquiescence variance is to select items strongly associated with the construct one is interested in measuring (Jackson, 1967d).

A final study serves to illustrate the potentially devastating effect that acquiescence may have in personality assessment by focusing directly upon an issue which is often of great interest: validity. Can acquiescence mask the existence of validity for personality items or scales? Bentler and Marshall[5] asked each member of a pair of friends to describe their own personality and to describe their friend's personality under one of two conditions. In the single-adjective condition, subjects described themselves or their friend by saying "true" or "false" to 300 adjectives designed to assess the three semantic differential dimensions as well as other personality attributes. In the polar-opposite condition, subjects described themselves or their friend by choosing one of two members of a polar-opposite pair of adjectives. The polar-opposite adjective list was identical, item by item, with the single-adjective list, save that a polar opposite was added to each adjective. Thus the subject had to choose one member of two positively worded, polar-opposite adjectives, rather than choose whether or not an adjective was descriptive.

Two sets of 100 pairs of subjects participated in the study, 100 pairs in each condition. Then it was possible to calculate 300 item-validity coefficients in each condition. Item validity was represented by the correlation between self and friend description, calculated across subjects. A high correlation would indicate, for example, that whenever a person described himself "active," his friend tended to describe him the same way. A zero correlation would indicate that, for this item, it was not possible to predict how a friend would describe a person from knowing the person's self-description. The 300 validity coefficients for each of the two conditions represent the basic data of interest. Is there greater validity, on the average, using one presentation format or the other?

From the discussion presented previously, it is obvious that the single-adjective format lends itself readily to the operation of both agreement and acceptance acquiescence. If these types of acquiescence detract from validity, as could be expected, the validity coefficients would be lower than they might otherwise be without the operation of acquiescence. The polar-opposite format, in contrast, essentially eliminates the effect of acceptance acquiescence since the subject must choose one positively phrased adjective from a pair. This format, however, may not eliminate the effect of agreement acquiescence, if we consider a tendency to agree with the first alternative as agreement acquiescence. Thus, the two conditions differ in that the single-adjective format has operative two acquiescence processes, the polar format only one. Conse-

---

[5] Bentler, P. M., and Marshall, M. Effect of presentation format on item validity. Unpublished manuscript, University of California, Los Angeles, 1969.

TABLE 9

Frequencies of Item Validities with Two Formats

Validity coefficient	Numbers of coefficients	
	Polar opposite	Single adjective
.72–.84	1	0
.60–.72	1	2
.48–.60	4	0
.36–.48	22	4
.24–.36	76	19
.12–.24	122	90
.00–.12	64	137
Below .00	10	48

quently, we would expect the validities to be higher in the polar format. The frequency distribution of validities presented in Table 9 speaks for itself. Validities are higher by far in the polar-opposite format than the single-adjective format. Eliminating acceptance acquiescence has markedly improved the predictive power of the adjectives.

We have reported this study particularly because we know that a content protagonist could well claim that the increased validities stem from the more clear-cut presentation of the content in the polar opposite condition: indicating that "disagreeable" is the opposite of "agreeable" makes it clearer to the subject just what "agreeable" means. It can be claimed that this increase in clarity is responsible for the higher validities. Indeed, since we are ourselves deeply concerned with developing assessment strategies which capture the content of personality, we too would endorse this technique from the point of view of content alone. We thus agree that content clarity contributed to the higher validities—indeed, we can envision studies aimed at identifying the *relative* magnitude of content and acquiescence effects in this situation. But for the present we must ask, Would a protagonist for the content position alone have carried out this study in the first place? And what would a content protagonist predict to happen to the interrelations among the various single adjectives when the total number of adjectives checked is taken into account, as in Bentler's (1969) study? If some content protagonists could be persuaded to investigate effects such as the ones described herein and to become interested in the many and varied influences which affect the identifiability of personality assessment content, we would be very pleased. After all, one of the reasons response styles are of concern is to provide a foundation for effectively tracking down the elusive "content." At the moment, however, it seems to us that a viewpoint that concerns itself with both content and response-style variance is more effective in generating research on basic processes involved in responses to structured personality inventories than is one exclusively oriented toward identifying content. In fact, it would appear that a content viewpoint will even have to struggle to succeed in explaining the data we have reviewed here.

REFERENCES

Anderson, N. H. Likeableness ratings of 555 personality-trait words. *Journal of Abnormal and Social Psychology*, 1968, **9**, 272–279.

Bentler, P. M. Review of J. Block, The challenge of response sets. *American Scientist*, 1966, **54**, 495A–496A.

Bentler, P. M. A new matrix for the assessment of factor contributions. *Multivariate Behavioral Research*, 1968, **3**, 489–494.

Bentler, P. M. Semantic space is (approximately) bipolar. *Journal of Psychology*, 1969, **71**, 33–40.

Bentler, P. M. A comparison of monotonicity analysis with factor analysis. *Educational and Psychological Measurement*, 1970, **30**, 241–250.

Bentler, P. M. Clustran, a program for oblique transformation. *Behavioral Science*, 1971, **16**, 183–185.

Berg, I. A. The unimportance of test item content. In B. M. Bass & I. A. Berg (Eds.), *Objective approaches to personality assessment*. New York: Van Nostrand, 1959.

Block, J. *The challenge of response sets*. New York: Appleton-Century-Crofts, 1965.

Bock, R. D., Dicken, C., & Van Pelt, J. Methodological implications of content-acquiescence correlation in the MMPI. *Psychological Bulletin*, 1969, **71**, 127–139.

Boyd, J. E., & Jackson, D. N. An empirical evaluation of judgment and response methods in multivariate attitude scaling. *American Psychologist*, 1966, **21**, 718. (Abstract)

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validations by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, **56**, 81–105.

Campbell, D. T., Siegman, C. R., & Rees, M. B. Direction-of-wording effects in the relationships between scales. *Psychological Bulletin*, 1967, **68**, 293–303.

Clayton, M. B., & Jackson, D. N. Equivalence range, acquiescence, and overgeneralization. *Educa-*

tional and Psychological Measurement, 1961, **21**, 371–382.

Couch, A., & Keniston, K. Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology*, 1960, **60**, 151–174.

Cronbach, L. J. Studies of acquiescence as a factor in the true-false test. *Journal of Educational Psychology*, 1942, **33**, 401–415.

Cronbach, L. J. Further evidence on response sets and test design. *Educational and Psychological Measurement*, 1950, **10**, 3–31.

Edwards, A. L. The social desirability variable: A review of the evidence. In I. A. Berg (Ed.), *Response set in personality assessment*. Chicago: Aldine, 1967.

Guttman, L. What lies ahead for factor analysis. *Educational and Psychological Measurement*, 1958, **18**, 497–515.

Jackson, D. N. *Personality Research Form Manual.* Goshen, N. Y.: Research Psychologists Press, 1967. (a)

Jackson, D. N. Review of J. Block, The challenge of response sets. *Educational and Psychological Measurement*, 1967, **27**, 207–219. (b)

Jackson, D. N. Balanced scales, item overlap, and the stables of Augeas. *Educational and Psychological Measurement*, 1967, **27**, 502–507. (c)

Jackson, D. N. Acquiescence response styles: Problems of identification and control. In I. A. Berg (Ed.), *Response set in personality assessment*. Chicago: Aldine, 1967. (d)

Jackson, D. N., & Lay, C. H. Homogeneous dimensions of personality scale content. *Multivariate Behavioral Research*, 1968, **3**, 321–338.

Jackson, D. N., & Messick, S. Content and style in personality assessment. *Psychological Bulletin*, 1958, **55**, 243–252.

Jackson, D. N., & Messick, S. Acquiescence and desirability as response determinants on the MMPI. *Educational and Psychological Measurement*, 1961, **21**, 771–790.

Jackson, D. N., & Messick, S. Response styles on the MMPI: A comparison of clinical and normal samples. *Journal of Abnormal and Social Psychology*, 1962, **65**, 285–299. (a)

Jackson, D. N., & Messick, S. Response styles and the assessment of psychopathology. In S. Messick & J. Ross (Eds.), *Measurement in personality and cognition*. New York: Wiley, 1962. (b)

Jackson, D. N., & Messick, S. Acquiescence: The nonvanishing variance component. *American Psychologist*, 1965, **20**, 498. (Abstract)

Janis, I. L., Mahl, G. F., Kagan, J., & Holt, R. R. *Personality*. New York: Harcourt, Brace, & World, 1969.

Lichtenstein, E., & Bryan, S. H. Acquiescence and the MMPI: An item reversal approach. *Journal of Abnormal Psychology*, 1965, **70**, 290–294.

Lubin, B., & Zuckerman, M. Affective and perceptual-cognitive pattern in sensitivity training groups. *Psychological Reports*, 1967, **21**, 365–376.

Messick, S. Response style and content measures from personality inventories. *Educational and Psychological Measurement*, 1962, **22**, 41–56.

Messick, S. The psychology of acquiescence: An interpretation of research evidence. In I. A. Berg (Ed.), *Response set in personality assessment*. Chicago: Aldine, 1967.

Messick, S. Response sets. In D. L. Sills (Ed.), *International encyclopedia of the social sciences*. Vol. 13. New York: Macmillan, 1968.

Mischel, W. *Personality and assessment*. New York: Wiley, 1968.

Morf, M. E., & Jackson, D. N. An analysis of two response styles: True responding and item endorsement. *Educational and Psychological Measurement*, in press.

Nunnally, J. C. *Psychometric theory*. New York: McGraw-Hill, 1967.

Osgood, C. G., Suci, G. J., & Tannenbaum, P. H. *The measurement of meaning*. Urbana: University of Illinois Press, 1957.

Peabody, D. Authoritarianism scales and response bias. *Psychological Bulletin*, 1966, **65**, 11–23.

Rorer, L. G. The great response-style myth. *Psychological Bulletin*, 1965, **63**, 129–156.

Rorer, L. G., & Goldberg, L. R. Acquiescence in the MMPI? *Educational and Psychological Measurement*, 1965, **25**, 801–817. (a)

Rorer, L. G., & Goldberg, L. R. Acquiescence and the vanishing variance component. *Journal of Applied Psychology*, 1965, **49**, 422–430. (b)

Rundquist, E. A. Item and response characteristics in attitude and personality measurement: A reaction to L. G. Rorer's "The great response-style myth." *Psychological Bulletin*, 1966, **66**, 166–177.

Trott, D. M., & Jackson, D. N. An experimental analysis of acquiescence. *Journal of Experimental Research in Personality*, 1967, **2**, 278–288.

Tucker, L. R. An interbattery method of factor analysis. *Psychometrika*, 1958, **23**, 111–136.

Wiggins, J. S. Substantive dimensions of self-report in the MMPI item pool. *Psychological Monographs*, 1966, **80**, (22, Whole No. 630).

(Received March 3, 1970)

Psychological Bulletin
1971, Vol. 76, No. 3, 205–210

ON FURTHER CONJECTURES REGARDING ACQUIESCENCE [1]

JACK BLOCK [2]

University of California, Berkeley

The two-dimensional interpretation of acquiescence offered by Bentler, Jackson, and Messick is evaluated. Although agreeing with Bentler et al. that their first dimension, agreement acquiescence (true-responding), is a relatively minor determiner of responses on inventories such as the MMPI, it is argued that their second dimension, acceptance acquiescence (the putative tendency to ascribe characteristics to oneself, regardless of the direction of item keying), is not likely to be of appreciable import for understanding the nature of responses to structured personality inventories. The relationships adduced by Bentler et al. as support for their view are evaluated, alternatively, as unresponsive to the evidential requirements of their position. Some recommendations for clarifying or responding to the possible role of acquiescence are offered.

The immediately preceding paper by Bentler, Jackson, and Messick (1971) seeks to renew interest in acquiescence as a possibly important determinant of responses to structured personality inventories. As reference to *Psychological Abstracts* will attest, interest in acquiescence has dropped away precipitously since the critiques advanced by Rorer (1965) and myself (Block, 1965). However, the newly offered conjectures and evidence advanced by Bentler et al. deserve consideration and response. Because of space limitations and also because my personal priorities at the present time are addressed to other scientific concerns, I shall comment only briefly and selectively on the views and analyses of Bentler et al.

1. In their introduction, Bentler et al. explicitly acknowledge the inadequacy of earlier conceptualizations of acquiescence employed by Jackson and Messick (1958). This inadequacy has been previously noted by diverse investigators; its acknowledgment by the protagonists of the earlier view closes a chapter in the history of personality assessment. A major consequence of this acknowledgment is that the studies earlier seeming to support this faulty conceptualization of acquiescence are now, by stipulation, of obscure import.

2. In defining the "content" and "style" positions, I believe Bentler et al. state the issues tendentiously. Not since the classic article by Meehl (1945) entitled "The Dynamics of 'Structured' Personality Tests" has the field of personality assessment viewed the content of inventory items in the naive, patent terms Bentler et al. attribute to the "content" position. To contrast, as Bentler et al. do, this unfairly defined "pure content position" with their all-encompassing "content-style position" is to construct a straw issue; certainly, it discourages serious discussion of the issues that may be involved.

A far better distinction between content and style was made by Jackson and Messick (1962) and is cited by Bentler et al. also. The reader is urged to note that where Bentler et al. construe content in face valid terms, Jackson and Messick earlier permitted item content to be "obvious or subtle . . . direct or indirect," so long as response consistency was achieved. In regard, then, to the conceptualization of "content" and of "style," 8 years have brought change but not progress, it would appear.

3. The new proposal of Bentler et al., that it may be useful to distinguish two types of acquiescence (agreement acquiescence and acceptance acquiescence), lacks the attractive simplicity of the earlier, admittedly inadequate conceptualization of acquiescence solely in terms of true-responding. It was only after reversal studies (e.g., Lichtenstein & Bryan, 1965; Rorer & Goldberg, 1965a, 1965b) demonstrated clearly that true-responding (agreement acquiescence) "must be a relatively minor determiner of responses on inventories such as the MMPI [Bentler et al., p. 191]," that acceptance acquiescence was invoked as a

---

[1] Preparation of this manuscript was facilitated by United States Public Health Service Research Grant MH-16080-01.

[2] Requests for reprints should be sent to Jack Block, Department of Psychology, University of California, Berkeley, California, 94720.

potential preservative of an acquiescence interpretation of structured personality inventories.[3] In these reversal studies, the investigators often did not try for what Bentler et al. call "polar-opposite" reversals because, for many inventory items, it is psychologically naive to posit a polar opposite. For example, what is the polar opposite of the item, "I feel lonely much of the time"? As another illustration, what is the polar opposite of the personality trait of "dominance"? Is it "submissiveness" or is it "egalitarianism"? [4]

The use of "negation reversals" avoids the trap of "polar reversals" which possibly are not received by respondents as psychological reversals. It was for this reason that later reversal studies tended to employ negation reversals, thereby excluding true-responding, regardless of content, as an explanation of an individual's response to inventory items. Whereupon Bentler et al. have conjectured, subsequent to these findings, that inventory respondents still are not attending to the meaning of the statements affirmed or rejected.

It is now proposed that individuals differentially respond to the presence or absence of negations in an inventory statement because of a differential tendency to ascribe characteristics to oneself, regardless of the direction of item keying. If the number of negations is zero (or even), the respondent high on acceptance acquiescence tends to say "true" to that item; if the number of negations is one (or odd), the respondent tends to say "false" to that item because the answer of "false" is a second negation that, in conjunction with the first negation, indicates a positive affirmation of the characteristic involved. The individual low on acceptance acquiescence tends to respond oppositely. In both instances, the respondent is presumed to be an algebraic purist, behaving with full rigor with respect to the number of negatives present while disregarding the meaning of the root verb or root adjective of the statement he is endorsing or rejecting.

Thus, for example, according to Bentler et al. a respondent high on the putative trait, acceptance acquiescence, would be expected to say "true" to both the statement "I am usually happy," and the statement "I am usually sad." He would say "false" to both the statement "It is not the case that I am usually happy," and the statement "It is not the case that I am usually sad."

Although a content interpretation of the MMPI item-reversal studies can be said to be confounded with certain implications of the acceptance acquiescence interpretation invoked post hoc by Bentler et al., other implications of their position would contradict the content position in provocative ways. Would subjects often claim being both *usually happy* and *usually sad*; both *usually a restless sleeper* and *usually a deep sleeper*; and so on?

I do not believe that human beings respond to personally relevant inventory statements in the automatic manner required by the acceptance acquiescence interpretation. The issue is, or can be, an empirical one. It would be a simple matter to construct quartets of items employing the format illustrated above with respect to the "happy" and "sad" characteristics. These items, suitably intermixed, could be administered to a suitable subject sample to see whether the particular response patterning required by the acceptance acquiescence conjecture is, in fact, to be seen reasonably often in the response patterns of subjects. This direct approach to evaluating the potency of the acceptance acquiescence interpretation seems to me to be far more compelling in regard to the kind of evidence it can provide than the indirect and confounded approach Bentler et al. have taken to adduce support for their view (cf. below).

4. The criticism by Bentler et al. of the MMPI is unproductive. Well-established points are scored again: certain MMPI scales

---

[3] Another reconceptualization of acquiescence, in terms of variance attributable to the true and the false portions of an inventory scale, also was offered about this time (Messick, S. Psychology and methodology of response styles. Paper delivered at Western Psychological Association, June 1965), but appears to have been abandoned—perhaps because this conceptualization lodges the response style of acquiescence in the inventory scale itself rather than in the responding individual (cf. Block, 1965).

[4] The empirical consequence of failure to recognize the frequently uncertain meaning of polar opposites is to be seen in Mogar's (1960) comparison of the Bass (1955) and Jackson and Messick (1957) reversed F Scales. These scales, independently constructed as polar opposites of the original F Scale, should have correlated highly with each other if polar opposites are easily and obviously achievable; instead, the correlation was only .35.

have poor psychometric properties, certain MMPI scales do not manifest the convergent and discriminant validity their labels require, the method of factor analysis applied to the usually employed MMPI scales indicates two factors "explain" much of the pattern of covariation of these scales.

But, of course, none of these remarks is positive evidence for the influence of acquiescence. To criticize the MMPI is easy; more difficult is educing evidence for the direct and unequivocal influence of acquiescence, of either type, in the MMPI. This Bentler et al. have not done. Indeed, careful reading of their section on the "Ubiquitous MMPI," reveals that nowhere is it asserted that acquiescence, of either kind, is one of the two main factors underlying MMPI scale intercorrelations. Although "desirability" is explicitly said to be one of the main factors, the other main factor is not claimed for acquiescence. It is therefore difficult to understand the intentions or the accomplishment of this section of the Bentler et al. argument.

5. The three lines of recent evidence that Bentler et al. have adumbrated for their new two-dimensional conceptualization of acquiescence I find disappointing.

The Bentler (1969) Study

This study employs the semantic differential procedure, thus shifting from the inventory domain wherein the issue of acquiescence was first raised and to which it is speculated the acquiescence interpretation would apply. Responding to the semantic differential procedure involves psychological demands on the subject that are very different from those created by an inventory like the MMPI. Thus, the applicability of any result achieved with the semantic differential cannot be generalized to the inventory medium without further research. Indeed, if one's intent is to apply an acquiescence interpretation to the inventory domain, it is an unwise detour to introduce a host of unknown variables, plus a significant generalization problem, by employing the semantic differential procedure.

The Bentler study, however, suffers from internal inadequacies as well. It recommits the old error of presuming psychologically polar opposites are easily attainable. Further, Bentler employs as an index of acquiescence a score that Bentler et al. admit confounds "in unknown combination" the two types of acquiescence they propose as important to distinguish. But this score, the total number of adjectives checked, does not confound only those two postulated kinds of acquiescence; in addition, this index is confounded as well with many other proclivities of the subject, for instance, the response set of extremeness, other response sets, and a host of personality variables, all operating in unknown but reliable combination.

Personality assessors over the years have come to recognize that scores such as the *number* of adjectives checked, the *number* of Rorschach responses emitted, the *length* of TAT protocols, and so on, although reliable, are not thereby unidimensional and unequivocal in meaning. Differences in weight are reliable also. But differences in weight may be due to differences in age, in sex, in genetics, in the number of amputations experienced, in nutrition, in time of day, in body fluid loss, in the wearing of clothes, and so on. To view weight as an index of limb loss (after all, amputees on the average weigh less than nonamputees) "in unknown combination" with clothes-wearing tendencies (after all, people who wear clothes, when weighed, weigh more, on the average, than people not wearing clothes) is to ignore the primary sources of variance underlying the weight index in its usual context of measurement. If large sources of the variance are unrecognized, or the size of the contribution of other variance sources cannot be assessed, the interpretation of an index in terms of small or unspecifiable variance sources severely misrepresents its meaning.

Positive evidence for interpretation in terms of acquiescence—of either or any kind—requires the use of direct and unconfounded measures, not simply a demonstration of the "existence of noncontent responding" using an index whose variance may be accounted for by myriad sources of large and unknown value. Phrased more vividly, a proof that someone is not in Berkeley is weak evidence for the assertion that he is at the North Pole.

Incidentally, the discovery by Bentler that removal of the influence of individual differences with respect to the total number of adjectives checked will result in content-meaningful relationships is not surprising. Personality assessors and psychometricians have long known of the intrusive, relationship-obscuring effect of such large, reliable, but variously based sources of variance. For example, Guilford's text (1954, Chapter 10) recounts the problem and illustrates (pp. 280–288) a variance-partitioning procedure for filtering out such obscuring static by simple adjustment techniques. As he noted, correlations that theoretically should be high may instead be low and even reversed if adjustments are not made for constant errors of rating or of checking. Although Guilford's discussion is phrased in terms of raters rating ratees rather than subjects checking concepts, his comprehensive analytical rationale applies directly. It fully and alternatively explains the phenomenon construed by Bentler et al. as evidence for the operation of acquiescence.

The Morf and Jackson Study

Many aspects of the Morf and Jackson (in press) study warrant comment, but only three concerns are noted here.

First, Morf and Jackson at the outset of their study sought to exclude from the arbitrary scales they were constituting all inventory items previously known to have high content saturations. Subsequently, acquiescence of both kinds was found.

This design decision astonishes me for it suggests that in order to "find" acquiescence, one must look for it under artificially constrained and irrelevant circumstances rather than in typical inventory domains where acquiescence first was sighted.[5] I am reminded of the drunk who, having lost his wallet in a dark alley, proceeded to look for it under a convenient street light rather than in the place where the wallet should be found.

Second, the inventory scales defined by Morf and Jackson, especially the facet-representing scales, impress me as noncoherent and as arbitrarily named. As only one of many possible examples, consider the items "I am not likely to take a watch apart just to discover how it works" and "I never seem to run out of questions on topics that interest me." Both of these items were assigned to the USNT scale, supposedly to reflect *an orientation toward understanding many areas of knowledge*, expressed via *self-description*, using *negative wording*, and keyed for a *true response*. The first of these items seems to me to reflect, if anything, a tendency opposite to that expressed by the second. Since no construct validation of these scales exists, the reader is asked to accept on faith the labels Morf and Jackson have provided for these scales. I am sure most psychologists will require greater justification of these scale names.

Third, the factor analysis on which Morf and Jackson rely for the claim that two kinds of acquiescence exist is technically faulty in a fundamental way. It will be recalled, from the equations of fundamental factor theory, that the reliability of a measure sets an upper bound to the communality or proportion of common-factor variance of that measure. For a given measure, communality cannot logically exceed reliability.

Reference to Table 6 of the Bentler et al. article reveals that the mean reliability of the 32 facet-representing six-item scales they employed was only .28 (i.e., on the average, only 28% of the variance of these crucially important scales was reliable). The reliabilities ranged from −.05 to .51. Further reference to their Table 7 shows that the mean communality of these same 32 scales was .39 (i.e., on the average, 39% of the variance of a scale was "explained" by the factor analysis). Of the communal variance, then, only 72% (.28/.39) can be said to be reliable. Especially embarrassing discrepancies abound (e.g., a communality of .45 when the reliability of the scale is only .10!).

I do not know what it means to extract random variance, then "interpreting" the resulting factors. I suspect Bentler et al. do not either. The "fit" that they see between their factor rotation of chancy variables and their hypothesized target matrix can well be illusory[6] in the light of the investigations by Horn (1967) and Humphreys, Ilgen, McGrath, and Montanelli (1969). These investigators factor analyzed random variables and have shown dramatically how complaisant factor analysis tends to be in fitting hypothesized factor matrices. The findings and cautionings of Horn and Humphreys et al. are compelling and have dismaying implications for factor analysis; they must be heeded by all who may be attracted to the method.

---

[5] Jackson and Lay (1968) formed inventory scales using the same item pool employed by Morf and Jackson, but did not exclude items having high content saturations. They found almost no acquiescence, of either kind.

Horn's summary is worth citing here:

> Seventy-four random variables for an N of 300 were generated [Morf and Jackson used an equivalent ratio of variables and subjects, 51 variables and 195 subjects], arbitrarily named, intercorrelated, factored, and rotated into a best (least-squares) approximation to an hypothesis factor matrix. The resulting factors were found to be quite interpretable and to have high hyperplane counts, although communalities and factor loadings were low. There was thus indication that if an investigator were willing to interpret relatively low loadings if these seemed to "make sense," he needn't bother to gather actual data: random variables may be labeled arbitrarily and pushed into solutions that make quite "good sense" [Horn, 1967, pp. 819–820].

The Bentler and Marshall Study[7]

This study demonstrates that a single-adjective format for description is less valid than a polar-opposite dimensional format that anchors the meaning of both ends of the dimension. Bentler et al. "know that a content protagonist could well claim that the increased validities stem from the more clear-cut presentation of the content in the polar-opposite condition . . . [and] agree that content clarity contributed to the higher validity." They ask, however, "would a protagonist for the content position alone have carried out this study in the first place?"

This protagonist would have to say "No" to this question, but for a reason different from the one Bentler et al. would prefer. He would have again remembered, for example, Guilford's chapter on rating scales (Guilford, 1954, Chapter 10), and known that greater discrimination and lesser equivocality result from anchoring both ends of a response continuum. Accordingly, greater validity will be found for a dimensional format as compared to a single-adjective format. Further demonstrational studies of the advantages of score unambiguity and of increased discrimination are matters of low priority in the field of personality assessment.

6. My remarks thus far have been focused, critically, on the Bentler et al. conceptualization and evidence. Some positive suggestions are in order as well.

Direct, unconfounded, construct-valid measures of each of the newly proposed aspects of acquiescence are required. This is a long and difficult responsibility, but it must be assumed if the two-dimensional interpretation of acquiescence is to become more than a post hoc conjecture. Further research employing admittedly confounded indexes will not be helpful.

It will not be fruitful to study acquiescence until the nature of the interdependency claimed by Bentler et al. to exist between the two kinds of acquiescence, and between content and each of the two kinds of acquiescence, is specified or shown to have lawful parameters. Presently, no conceivable result can embarrass this all-embracing interpretation because the principles of interdependency have not been stated prior to and independent of the relationships to be evaluated.

When direct and unconfounded measures of acquiescence have been developed from different domains, these measures should be correlated to establish that, indeed, acquiescence is a broad response disposition. This has only rarely been tried in the past, and, when tried, has offered only disappointing results.

In the effort to develop compelling evidence for acquiescence, further usage of the method of factor analysis as it has been employed in the past would appear to be unprofitable.

Until positive, systematic evidence, based upon construct-valid indexes, is developed attesting to the importance of agreement acquiescence and acceptance acquiescence, investigators have a choice of three alternatives. They can ignore the possible influence of acquiescence, usually with no particular danger. Or, and better, they can adjust their indexes via techniques such as the Guilford procedure or partial correlation to eliminate the relationship-obscuring effects of the specific variance introduced by diverse subjects. Or, and best, they can prevent at the outset the possible intrusion of acquiescence and other possible sources of relationship-obscuring variance by appropriate procedures (e.g., the forced-choice method). For most purposes of psychological assessment, it is not necessary to study why certain response tendencies obscure relationships—the reasons may be uninteresting or trivial or idiosyncratic ones. Rather, it may be quite sufficient simply to adjust or to prevent, by suitable methodological or procedural means, the obscuring effects of such response styles rather than become preoccupied with them.

---

[6] Morf (1969, pp. 73–74), in his original analyses (Morf, M. E. An analysis of two response sets: True responding and item endorsement. Unpublished doctoral dissertation, University of Western Ontario, 1969), could not find a factor representing acceptance acquiescence when he partitioned his subject sample into male and female subsamples, despite efforts to maximize fit to this hypothesis and despite the strong part–whole relationships between his subsamples and his total sample.

[7] Bentler, P. M. and Marshall, M. Effect of presentation format on item validity. Unpublished manuscript, University of California, Los Angeles, 1969.

REFERENCES

Bass, B. M. Authoritarianism or acquiescence. *Journal of Abnormal and Social Psychology*, 1955, **51**, 616–623.

Bentler, P. M. Semantic space is (approximately) bipolar. *Journal of Psychology*, 1969, **71**, 33–40.

Bentler, P. M., Jackson, D. N., & Messick, S. The identification of content and style: A two-dimensional interpretation of acquiescence. *Psychological Bulletin*, 1971, **76**, 186–204.

Block, J. *The challenge of response sets*. New York: Appleton-Century-Crofts, 1965.

Guilford, J. P. *Psychometric methods*. (2nd ed.) New York: McGraw-Hill, 1954.

Horn, J. L. On subjectivity in factor analysis. *Educational and Psychological Measurement*, 1967, **27**, 811–820.

Humphreys, L. G., Ilgen, D., McGrath, D., & Montanelli, R. Capitalization on chance in rotation of factors. *Educational and Psychological Measurement*, 1969, **29**, 259–272.

Jackson, D. N., & Lay, C. H. Homogeneous dimensions of personality scale content. *Multivariate Behavioral Research*, 1968, **3**, 321–337.

Jackson, D. N., & Messick, S. A note on ethnocentrism and acquiescent response sets. *Journal of Abnormal and Social Psychology*, 1957, **54**, 132–134.

Jackson, D. N., & Messick, S. Content and style in personality assessment. *Psychological Bulletin*, 1958, **55**, 243–252.

Jackson, D. N., & Messick, S. Response styles and the assessment of psychopathology. In S. Messick & J. Ross (Eds.), *Measurement in personality and cognition*. New York: Wiley, 1962.

Lichtenstein, E., & Bryan, S. H. Acquiescence and the MMPI: An item reversal approach. *Journal of Abnormal Psychology*, 1965, **70**, 290–294.

Meehl, P. E. The dynamics of "structured" personality tests. *Journal of Clinical Psychology*, 1945, **1**, 296–303.

Mogar, R. E. Three versions of the *F* Scale and performance on the semantic differential. *Journal of Abnormal and Social Psychology*, 1960, **60**, 262–265.

Morf, M. E., & Jackson, D. N. An analysis of two response styles: True responding and item endorsement. London, Canada: University of Western Ontario, 1969.

Rorer, L. G. The great response-style myth. *Psychological Bulletin*, 1965, **63**, 129–156.

Rorer, L. G., & Goldberg, L. R. Acquiescence in the MMPI? *Educational and Psychological Measurement*, 1965, **25**, 801–817. (a)

Rorer, L. G., & Goldberg, L. R. Acquiescence and the vanishing variance component. *Journal of Applied Psychology*, 1965, **49**, 422–430. (b)

(Received June 1, 1970)

Psychological Bulletin
1971, Vol. 76, No. 3, 211–214

A COEFFICIENT OF DIRECTIONAL CORRELATION FOR TIME-SERIES ANALYSES [1]

ROBERT F. STRAHAN [2]

University of Rochester

The tendency for two variables to change values in the same direction from one sampling occasion to the next is termed *directional correlation*. The *coefficient of directional correlation* (*d coefficient*) will often parallel other measures of bivariate association, especially rank correlation. In time-series (*P*-type) situations, however, directional correlation may diverge greatly. Although modifications of traditional correlational statistics will suffice in some of these situations to reflect the directional correlation present, they will not be generally adequate; moreover, the modifications involve considerable effort. The easily computed *d* coefficient is recommended when directional correlation is a phenomenon of interest.

The association between two variables sampled over time from the same subject, Cattell's (1952) *P*-type covariation, can be described statistically by one or another correlation coefficient. One kind of association may be encountered, however, that is not always reflected by usual correlational techniques. The term *directional correlation* is given to the tendency for two variables to change values in the same direction from one sampling occasion to the next. If, for example, increases in variable X are accompanied for the most part by increases in variable Y, and decreases in X by decreases in Y, then one can speak of some degree of positive directional correlation between variables X and Y.

In many, possibly most, time-series situations, directional correlation will parallel rather closely other kinds of association, such as product-moment and, particularly, rank correlation. When this is not the case, a measure of directional correlation may be of value. Such a measure is easily derived.

THE *d* COEFFICIENT

Suppose that values of X and Y are recorded on n successive occasions. Comparison is to be made of the (X,Y) values of each sampling occasion after the first with the (X,Y) values of the immediately preceding occasion. A plus score is given for each instance in which the direction of X is the same as that of Y (both variables increase in value or both decrease), a minus score for each instance in which the direction of X is opposite that of Y (one variable increases, the other decreases), and zero or no score for each instance of tied data (either or both variables remain unchanged). The *coefficient of directional correlation* (*directional* or *d coefficient*) is then defined as

$$d = \frac{p - m}{p + m} \qquad [1]$$

where p is the number of pluses, and m the number of minuses. If there are no tied data, $p + m = n - 1$, and d can equivalently be written as

$$d = \frac{p - m}{n - 1}. \qquad [2]$$

It is evident that Equation 1 in effect omits tied data from statistical consideration. In situations where such omission is not desired, Equation 2 can be used.

As do most other correlational statistics, d ranges from -1 through 0 to $+1$, these values reflecting, in turn, perfect negative directional correlation, the absence of directional correlation, and perfect positive directional correlation.

The following is an illustration of the computation of d. The six bivariate observations are listed from top to bottom in order of sampling.

---

[1] This work was initiated in the psychophysiological laboratory of D. T. Lykken, University of Minnesota, and was in part supported by National Institute of Mental Health predoctoral fellowship 5 T1 MH-6964-04 and a North Atlantic Treaty Organization postdoctoral fellowship in science. Among those providing helpful comments were R. D. Miller, C. G. Katzenmeyer, C. K. Strahan, and D. T. Lykken.

[2] Requests for reprints should be sent to Robert Strahan, Department of Psychology, University of Rochester, Rochester, New York 14627.

X	Y	
6	4	
8	5	+
2	3	+
7	1	−
7	2	0
9	5	+

$$d = \frac{p-m}{p+m} = \frac{3-1}{3+1} = 0.50.$$

When the various interrelations among several variables are sought, a simple device makes possible greater efficiency in calculation. That device consists in first scoring pluses, minuses, and zeroes for each variable separately, as that variable increases, decreases, and remains constant in value, then calculating the various ds on the basis of these preliminary scores.

SOME STATISTICAL RELATIONS

There are close associations between the directional coefficient and certain other statistics. d is formally equivalent to Du Mas' (1946) coefficient of profile similarity, r_{ps}, though the two statistics were developed independently for different purposes. d is also related to τ, Kendall's (1962) coefficient of rank correlation. τ can be expressed in such form that, in the case of untied data, a parallel is seen immediately:

$$d = \frac{p-m}{n-1}, \quad \tau = \frac{P-M}{\frac{n(n-1)}{2}} \qquad [3]$$

where p and P are the number of pluses, and m and M the number of minuses scored for the sample of size n. Scoring for τ, however, involves all possible pairings of the n (X, Y) values in the sample ($n[n-1]/2$ in number), whereas scoring for d involves only a temporally ordered subset ($n-1$ in number).

In the general case of tied data, d is directly analogous to γ, Goodman and Kruskal's (1954; see also Hays, 1963) variant of τ, rather than to Kendall's tied-data expression:

$$d = \frac{p-m}{p+m}, \quad \gamma = \frac{P-M}{P+M},$$

$$\tau = \frac{P-M}{\sqrt{\left[\frac{n(n-1)}{2} - T\right]\left[\frac{n(n-1)}{2} - U\right]}}. \qquad [4]$$

(It should be noted that Kendall recognizes tied-data situations in which τ as given in Equation 3, rather than Equation 4, is the appropriate statistic.)

A more fundamental relation between directional and rank correlation can be stated. When sampling randomly from a given untied population, τ is an unbiased estimator of the population rank-correlation coefficient, τ'. Under these same conditions, d also is an unbiased estimator of τ'.[3] τ extracts more information from the sample and is consequently the more efficient statistic. d, on the other hand, is considerably easier to calculate, increasingly so as n becomes larger.[4]

TIME-SERIES CASE

The likeness of d and τ (in its various forms) has been shown in the previous section. The difference giving point to this paper arises, as noted earlier, in the time-series context, where sampling is temporal rather than random, and observations generally dependent rather than independent. d and τ might in this case still be comparable in value, but they might also be quite divergent. It is possible to observe strong directional correlation between variables

---

[3] That d is an unbiased estimator of τ' can perhaps be seen directly. A proof requires that the mean of d over all possible random samples of size n be equal to τ'. It can be shown that each of the

$$\frac{N(N-1)}{2}\ (X,Y)$$

pairs in the population of size N is scored in

$$\frac{2(n-1)(N-2)!}{(N-n)!}$$

of the

$$\frac{N!}{(N-n)!}$$

possible random samples. The mean of d can then be expressed as

$$\frac{P-M}{n-1}\left[\frac{2(n-1)(N-2)!}{(N-n)!} \Big/ \frac{N!}{(N-n)!}\right]$$

which reduces to

$$\frac{P-M}{\frac{N(N-1)}{2}} = \tau'.$$

[4] A pertinent reference, discovered after completion of this manuscript, is the paper by Moore and Wallis (1943). That work deals in part with what is essentially directional correlation.

whose rank correlation is negligible or even opposite in sign, and vice versa. The source of the divergence resides in the fact that d tends to reflect short-term (phasic) association to the relative neglect of the long-term (tonic) association present, whereas the reverse holds for τ. A and B of Figure 1 are extreme examples constructed (with no ties, as is true of subsequent examples) to illustrate this differential sensitivity to these two components of covariation. Sampling points for the two variables (X and Y) are represented by closed circles. C of Figure 1 is an example of d–τ disparity caused primarily by a nonmonotonic association with time of one of the variables. D illustrates yet another kind of time-series situation in which directional and rank correlation differ markedly.

The significance of the phasic-tonic distinction is further illustrated by considering the effect on the examples given of decreasing sampling frequency by a factor of two or more. Most dramatic is the persisting reversal of d's sign in A and B. d is now reflecting (even more strongly than τ, as it happens) the strong tonic relations between variables. Insofar as the kind of association revealed depends on sampling frequency, that factor is clearly of concern to the investigator. (Also at issue in some situations is whether sampling is to be strictly temporal—for instance, every 10 seconds—or to bear some relation to whatever stimuli may be present.)

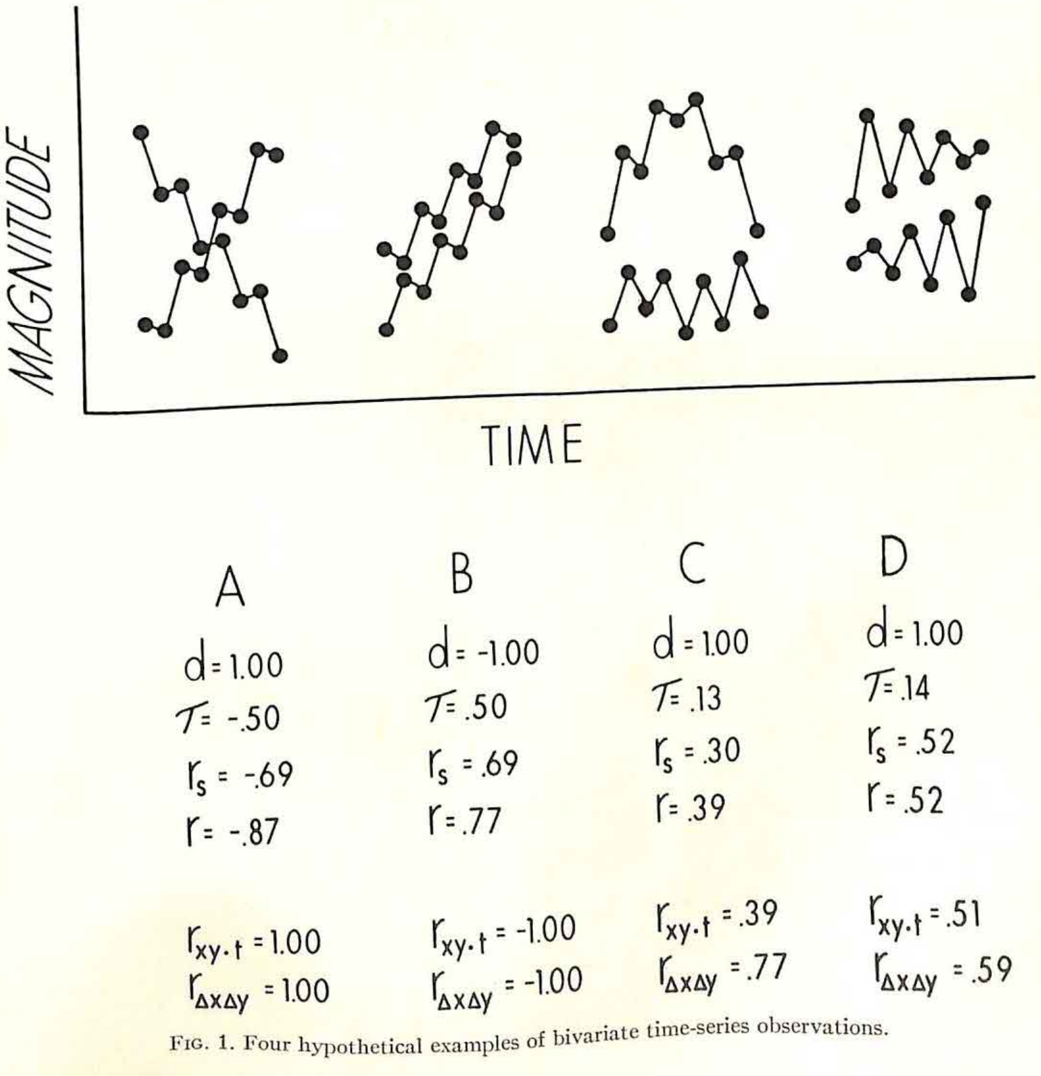


FIG. 1. Four hypothetical examples of bivariate time-series observations.

Although it is obvious in the examples given that none of three common correlational statistics—Pearson's r, Spearman's rank-correlation coefficient, r_s, and Kendall's τ—is reflecting well the directional correlation present, it might be suspected that some modification of these traditional statistics, such as partial correlation, would serve as an approximation. Two such modifications are discussed. Their illustration is in terms of r, however, rather than τ. Notwithstanding the fact that d is more closely related to τ than to r, the particular examples used are more instructive here in terms of product-moment correlation.

In certain instances, but not always, directional correlation will be approximated by rank or product-moment correlation with time as a partialed-out third variable. The instances are those in which associations between time and the basic variables are substantially linear (r) or monotonic (τ). (There is no partial counterpart of r_s.) A and B of Figure 1 are examples in which partial correlation technique is quite effective (due to the strong linear associations with time); C and D are examples in which partialing is ineffective (due to the absence of linear—even monotonic—associations).

The second approach to approximation of directional correlation involves change scores. Here the variables are $\Delta X_i = X_{i+1} - X_i$ and $\Delta Y_i = Y_{i+1} - Y_i$, rather than X_i and Y_i (where i, the order of sampling, ranges from 1 to $n - 1$). Interestingly, d is a special case of this approach in which only the signs of the change scores enter into computation. If ranks of the change scores are utilized, a τ or r_s coefficient results, and if magnitudes as well enter in, one has an r coefficient. Even here, however, directional correlation may not always be reflected by any statistic other than d. In A and B of Figure 1 there is the strongest congruence between r and d, but in C and D, only partial congruence—especially in D. The reason for this latter deviation is apparent: Although X and Y change values consistently in the same direction, the magnitudes of changes are inversely related.

The consequence of these remarks is that if directional correlation is of interest, one is assured an adequate measure only by using the d coefficient. Moreover, that coefficient is far easier to compute than any alternative statistic. (A d-like index can be calculated in phi coefficient fashion using the same data—number of plus and minus scores—utilized by d proper. Although yielding generally similar values, this alternative statistic involves considerably more computational effort and does not appear to have any compensating advantage.)

Two concluding comments are in order. The first is reiteration that the directional coefficient is presented and discussed as a descriptive statistic. (On the inferential complications stemming from the general dependence of time-series observations see, for instance, Cattell, 1963, and Holtzman, 1963). The second is the caveat that this paper makes no brief for the general utility of the directional coefficient. Although the statistic has had some application in the psychophysiological context in which it was developed, it remains to be seen the extent to which directional correlation is distinct from other kinds of relation between time-series variables, and the degree to which that distinction is important to investigators.

REFERENCES

CATTELL, R. B. The three basic factor-analytic research designs—Their interrelations and derivatives. *Psychological Bulletin*, 1952, **49**, 499–520.

CATTELL, R. B. The structuring of change by P-technique and incremental *R*-technique. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press, 1963.

DUMAS, F. M. A quick method of analyzing the similarity of profiles. *Journal of Clinical Psychology*, 1946, **2**, 80–83.

GOODMAN, L. A., & KRUSKAL, W. H. Measures of association for cross classifications. *Journal of the American Statistical Association*, 1954, **49**, 732–764.

HAYS, W. L. *Statistics*. New York: Holt, Rinehart & Winston, 1963.

HOLTZMAN, W. H. Statistical models for the study of change in the single case. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press, 1963.

KENDALL, M. G. *Rank correlation methods*. (3rd ed.) New York: Hafner, 1962.

MOORE, G. H., & WALLIS, W. A. Time-series significance tests based on signs of differences. *Journal of the American Statistical Association*, 1943, **38**, 153–164.

(Received March 5, 1970)

Psychological Bulletin
1971, Vol. 76, No. 3, 215–230

INTERCULTURAL COMMUNICATION AND THE ADJUSTMENT OF THE SOJOURNER

MICHAEL BREIN
University of Hawaii

AND

KENNETH H. DAVID[1]
Peace Corps Training Center, University of Hawaii

Intercultural adjustment is described, in general, by means of a W-curve function or culture shock. However, very little apparent integration exists among the various approaches for explaining or understanding the sojourner's adjustment. Social patterns of behavior and personality traits have not been very successful in explaining intercultural adjustment, while social interaction and background and situational factors have been of somewhat greater value. These divergent approaches can be more meaningfully understood in relation to adjustment to the degree that they facilitate or impede intercultural communication. Thus, intercultural communication appears to be both an integrating and crucial factor for understanding the adjustment of the sojourner.

The purpose of this article is to discuss social psychological factors related to the adjustment problems of the sojourner. Books and articles by Cleveland, Mangone, and Adams (1960), Ekroth (1968), Gardner (1962), Gullahorn and Gullahorn (1963), Hall (1959), Hall and Whyte (1963), Jacobson (1963), Kelman (1965), and Major[2] have discussed various aspects of sojourning and adjustment.

The term *sojourner* refers to many types of travelers including students, trainees, technical assistants, tourists, businessmen, military personnel, missionaries, foreign service officers, professors, and others. The sojourn research has primarily considered the foreign national who has sojourned in the United States, although there appears to be an increasing interest in the American who has traveled to other countries (e.g., Major, 1965, see Footnote 2; Smith, Fawcett, Ezekial, & Roth, 1963). Moreover, among those studies that have dealt with the American sojourner, there has been a disproportionate emphasis on the study of students and educational advisors. According to Patterson (1969) and the United States Bureau of the Census (1969), pleasure and personal reasons were indicated by approximately 78% of all United States passport recipients in 1967 and 1968 as the two main purposes for their travel. Yet, these individuals have been relatively neglected by the sojourn research. Another emphasis of this kind of research has been on the study of the traveler during the period of his stay in the foreign or *host* culture. There have been relatively few studies of the adjustment of the returnee to his home country, the permanence of the effects of the sojourn, the effects of the sojourner on his hosts, and the effects of the returnee on his own country.

Because of the increased opportunities for international travel, more and more persons from many countries are interacting with one another. With such increased frequencies of intercultural interaction, the sojourner is faced with a number and variety of adjustment difficulties which can be traced to both the obvious and more subtle differences between cultures. The major purpose of this article, therefore, is to discuss the social psychological aspects of the adjustment of the sojourner and to provide an integration of the research concerned with this topic.

Adjustment problems can have serious consequences for the sojourner. Thomson and English (1964), for example, reported that adjustment problems constituted over 60% of all prematurely returning Peace Corps Volunteers prior to 1963. The problems of adjustment are also seen to range widely in extent, focusing on the sojourner's relationship to himself, to others, to the environment, and to his activities within the host culture. Smith (1955), for example, discussed the adjustment difficulties that foreign students have in the United States in relation to communication, learning the ways of academic life in the United States, maintaining self-esteem, gaining acceptance and making friends, and balancing group memberships and loyalties. And Dicken (1969), Fisher, Epstein, and Harris (1967), Mischel (1965), and Smith et al. (1963) concerned themselves with the adjustment of Peace Corps Volunteers in relation to their jobs, their physical well-being, and their in-country locations.

---

[1] Requests for reprints should be sent to Kenneth H. David, P. O. Box 856, Peace Corps Training Center, University of Hawaii, Hilo, Hawaii 96720.

[2] Major, R. T., Jr. A review of the research on international exchange. Unpublished manuscript, Putney, Vermont: The Experiment in International Living, 1965.

A major problem in understanding the topic of the adjustment of the sojourner is that the approaches of various investigators have been so divergent that it is difficult to either interrelate their findings or to develop any consistencies among the factors deemed relevant to intercultural adjustment. Moreover, even when investigators apparently study the same or similar factors, the results usually fail to show any consistent patterns. Several of the more common approaches to describing and explaining the sojourner's adjustment include the following: curves of adjustment, culture shock, personality typologies and traits, background and situational factors, and social interaction. These approaches have varied in their usefulness in explaining the process of intercultural adjustment and are discussed in detail in later sections of this paper.

An assumption is made by the authors that the sojourner's successful adjustment to an intercultural experience is highly dependent on his achievement of effective interpersonal relations with his hosts. A necessary prerequisite to establishing such effective social relationships, however, is the development of understanding between the host and visitor. To this end, the effective exchange of information, that is, communication, on both verbal and nonverbal levels of behavior, is considered by the authors as being crucial to the development of such understanding. One mode of integrating the seemingly rather divergent approaches to understanding the adjustment process, therefore, is to interpret adjustment within the broad framework of intercultural communication as outlined above. In this manner, a variety of factors may be seen to influence the adjustment process insofar as they affect the process of information exchange or communication.

U AND W CURVES OR ADJUSTMENT FUNCTIONS

The sojourn literature indicates that the adjustment of the traveler seems to follow a predictable temporal pattern. That is, there seems to be a curvilinear relationship between adjustment and time sequence of the sojourn. Lysgaard (1955), in a study of Scandinavian Fulbright Grantees studying in the United States, first described this curvilinear function and referred to it as the U-curve phenomenon. And Sewell and Davidsen (1956) reported a U-curve function of academic personal adjustment for Scandinavian students visiting the United States. Initially, researchers reported only a U curve, because the studies were limited to the sojourner's adjustment within the host culture; however, a more recent development suggests that a W-curve function may be a more comprehensive description of the adjustment process. The W curve is essentially an extension of the single-U curve to a double-U curve (UU curve), that is, a W curve. The W curve of adjustment has been described by Gullahorn and Gullahorn (1963, 1966). It encompasses both the sojourner's adaptation to a foreign culture and his readjustment to his home culture. Thus, the W-curve function represents the adjustment of the sojourner along a temporal dimension. In a very general manner, the sojourner tends to undergo a decline in adjustment shortly after entering a foreign culture, which is followed by a recovery stage with a resultant increase in adjustment; then, on returning home, the sojourner undergoes another decrease in adjustment followed by a second stage of recovery. The degree and duration of the adjustment decline for an individual sojourner would depend on a number of variables, which are discussed in the later sections of this article. Thus, the W curve may be descriptive of the change in adjustment of a larger number of sojourners, although there may be a great deal of variability among individual curves of adjustment. In some instances a sojourner may never quite fully recover from his initial difficulties, whereas others may experience little more than a mild annoyance with no appreciable decrease in adjustment. Also, some sojourners may encounter few adjustment difficulties upon returning home, while others may undergo a rather severe reentry crisis.

During the initial phase of the W-curve phenomenon, which may last for several months, depending on the length of the sojourn, the sojourner shows a relatively satisfactory adjustment to the host country. The visitor is mainly a spectator, absorbing the sights, forming impressions, and expressing enthusiasm. His interactions with the host nationals are quite limited and uninvolved. Tourists or short-term sojourners are usually characterized as rarely advancing from this superficial level of interaction with host nationals, perhaps because their purposes are related to recreation, relaxation, and sightseeing, which may preclude any serious involvements with the hosts. A decrease in the sojourner's adjustment, that is, a downward slope in the W curve, begins after this initial spectator phase. The sojourner has been in the host country for a period of time and encounters gradually increasing demands to adjust to a new and often strange environment. The sojourner's previously limited involvements with the hosts no longer suffice to satisfy his needs to cope with his intercultural experience. He attempts to unravel the complexities of the culture, and to accomplish this he develops the need to establish more enduring social relationships with the hosts. He soon learns that his elementary knowledge of the host language is no longer adequate enough to cope with the increasing complexities of social communication. This is a time when the sojourner's involvement with the host culture becomes critical; he is at the point where he is in most need of satisfactory relationships with others, yet, he lacks the social tools or means to bring these about. This is the low point of the sojourn where the characteristic symptoms of culture shock occur, as described by Oberg (1960) and others. After this, a gradual recovery in the sojourner's adjustment occurs (an upward trend of the W curve), where he begins to become aware of the more subtle cues of the host culture and begins to develop a fluency in the language. He develops a strategy for working through various intercultural difficulties. When the sojourner returns to his home culture he may undergo what has been termed a "reentry crisis," which is analogous to the first drop in the W curve. His decline in adjustment, however, seems to be somewhat less in intensity than during his initial culture shock reaction to the host culture.

The sojourn literature indicates that the W curve may show cultural variation. For example, Pool (1965) reported that among foreign students to the United States, those coming from countries that are relatively more similar to the United States, that is, other Western nations, differ in their adjustment patterns from those students from countries less similar to the United States, that is, Asian or non-Western nations. Sewell and Davidsen (1956) reported that language facility, social interaction with Americans, previous contact with other cultures, and personality characteristics may influence the rate that a sojourner to the United States passes through the various phases of the W curve.

CULTURE SHOCK

Rather than describing adjustment strictly in terms of curve functions per se, other researchers have portrayed adjustment in verbal terms as a series of sequential stages. Although differing with respect to numbers and kinds of stages, culture shock, role shock, language shock, culture fatigue, and reentry crisis all bear in common a close resemblance to one or more aspects of the U and W phenomena.

Hall (1959) defined culture shock as "a removal or distortion of many of the familiar cues one encounters at home and the substitution for them of other cues which are strange [p. 156]." Oberg (1960) is generally credited for having introduced the term in one of his earlier articles. He conceived of culture shock as an "occupational disease" of people who suddenly find themselves located in a culture very different from their own. Oberg stated that culture shock is precipitated by the anxiety that results from losing one's cues to social interaction. The cues include the many ways in which one orients himself to daily life: when to shake hands, when and how much to tip, when to accept invitations, when to take statements seriously, etc. The cues, which may be customs, gestures, facial expressions, or words, are primarily unconscious and are as much a part of one's culture as the language.

Oberg described several stages of culture shock: (*a*) Stage 1 is a period of incubation, during which time the sojourner may feel highly elated. He may be living in a good hotel where conditions are comfortable, the sights are intriguing, the hosts are courteous, and his expectations for the future are positive. During this period he may find a house, maids, schools for his children, etc. (*b*) Stage 2 is a period of crisis resulting from the genuine difficulties that the sojourner may begin to encounter in a different culture. The activities of daily living that had been previously taken for granted become insurmountable problems. He may associate with other sojourners and criticize the host country and its people. Thus, foreign nationals may be described in derogatory terms as a result of the sojourner's attempt to rationalize his adjustment problems. (*c*) Stage 3 is a period of recovery in which the sojourner begins to understand some of the cues of the host culture. He regains his sense of humor, and he may even imagine himself an authority on the host culture. (*d*) Stage 4 is the complete or nearly complete period of recovery in which the sojourner accepts the host culture. Although he may not be overly enthusiastic about the country, he is at least able to enjoy his experiences. (*e*) Stage 5 is a period that occurs when the sojourner returns home. It is a culture shock in reverse, but one that seems to be less serious than the original shock. Aside from the "disease" model that Oberg used to describe culture shock, Hall and Oberg appear to be in basic agreement. Gullahorn and Gullahorn (1963) equated culture shock to the cognitive dissonance that results from the discrepancy between the sojourner's expectations of the host culture and his observations of that culture. Also, cognitive dissonance may occur as the result of the sojourner's participation in role behaviors that are at variance with his own values. Smalley (1963) proposed four stages of culture shock: (*a*) A stage of fascination in which there are various barriers preventing social interaction with host nationals. (*b*) A stage of hostility which occurs as the permanency of the residence develops. There is more social interaction within the host country, but the sojourner emphasizes the superiority of his own culture. (*c*) A stage of improved adjustment, with an expression of humor and lessening of tensions. (*d*) A stage of biculturalism, in which the sojourner develops an understanding of the host culture and acts in accordance with its norms. These stages appear to be similar to those proposed by Oberg.

The crises of Peace Corps Volunteers have been described by Pearson (1964). During the first crisis the volunteer is compelled to explain that he is "none of these things." He believes that he must explain his purpose to his hosts and dispel their misconceptions. The second crisis is one of acceptance of the situation; the romance and adventure are gone; the volunteer becomes bored, lacks intellectual stimulation, and his work becomes tedious. He begins to read more and discovers a "society of solitude." The third crisis is one of self-awareness and few frustrations. There is a feeling of identity and an awareness of the "different life."

According to Foster (1965), few first-time longer-term sojourners recover from the culture-shock experience in less than 6 months, and it is not uncommon for it to last for a year, although a resilient sojourner may recover within 3 months. Foster suggested that one culture-shock experience will not immunize one against future shocks, although future experiences may be less severe, and the readjustment stages somewhat shortened.

Smalley (1963) considered language to be the primary determinant of culture shock, because of the importance of language as a medium of communication. He added that the study of language per se may result in an acute shock for some people. Smalley appears to disagree with Hall over the relative importance of different types of communication, with Smalley emphasizing verbal communication, and Hall stressing the nonverbal aspects.

In an article about overseas American technical assistants, Byrnes (1966) described "role shock" as being somewhat distinct from culture shock. Unlike culture shock, role shock was reported to be more severe, to increase over time, to frequently reach its peak at the midpoint of the expected length of the sojourn, and to rarely result in complete recovery. Role shock for technical assistants was believed to result from ambiguity in their professional roles, unsuccessful relationships with their host-country counterparts, and problems with communication and participation in the host-country hierarchy.

Guthrie (1966) suggested that the term "culture fatigue" may be a more appropriate description of the reactions of Peace Corps Volunteers to the Philippines. The hospitality and deference of the Filipino, in conjunction with the fairly common use of English and the awareness of certain Western values, may result in the American sojourner believing that the Philippine culture is very similar to that of his own. However, in spite of these apparent similarities as well as the adoption of some Spanish and American customs by the Filipinos, there remain many elements that are unique to that country. The apparent cultural similarities may result in the sojourner experiencing confusion and annoyance, and when these frustrations endure over a period of time they may be termed "culture fatigue." Szanton (1966), too, discussed culture fatigue among Peace Corps Volunteers in the Philippines. He considered culture fatigue to be the exhaustion that results from the endless number of minute adjustments required for long-term living in a foreign culture.

On returning to his home country, the sojourner may undergo what has been termed a reentry crisis. The crisis of the returning Peace Corps Volunteer has been described by Pearson (1964) and Stolley (1965). The volunteer, having lived for 2 years in a developing country, may find difficult his readjustment to the relative affluence of the United States and the seemingly superficial values espoused by the mass media. Moreover, he may have difficulties in adjusting to a change in roles. As a volunteer he may have assumed many responsibilities, but, after a while at home, may find himself in a relatively ambiguous situation with minimal responsibilities and resultant feelings of disillusionment and aimlessness. The returning volunteer is confronted with many behaviors that prior to his departure from the United States were primarily unconscious, unnoticed, unquestioned, and taken for granted. But on returning home and becoming aware of these previously accepted behaviors, these may no longer seem meaningful. The apparent values as reflected in the heavy traffic, ubiquitous television, rushing about, etc., may be very difficult for the returnee to reconcile with the values that he acquired while in the host country, for instance, values of social concern which he acquired as a result of teaching, preventing illnesses, or increasing food production.

SOCIAL PATTERNS, PERSONALITY, AND ADJUSTMENT

Various investigators have attempted to construct sets of characteristics and distinct social patterns that could be used to distinguish between groups of sojourners. There may be a limited number of alternatives that a sojourner may choose from to cope with a foreign environment, and each alternative would include a characteristic modus operandi. Once the different social response patterns have been determined, a fairly accurate description of the social behavior of any sojourner, after he adopts one of the characteristic patterns, is possible. For example, Shelton (1964) observed a characteristic pattern of social interaction, which he termed the "Miss Ophelia Syndrome," among a group of investigators conducting field research in Africa. The syndrome consisted of an avoidance of physical contact with the Africans and avoidance of accepting Kola (nuts), food, or wine from them. The usual reason, or rationalization as Shelton perceived it, was that the investigators were acting in a scientific manner of detachment which was appropriate for conducting their research. However, the Africans, as well as other host nationals, are likely to have interpreted the aloofness of the researchers in a manner that may have jeopardized the probability of the researchers' success. The villagers may have believed, for example, that the "Miss Ophelias" considered themselves to be superior to the Africans.

Sewell and Davidsen (1956) described four kinds of sojourners and the characteristic patterns of adjustment that they adopt: (*a*) Detached Observers, (*b*) Promoters, (*c*) Enthusiastic Participants, and (*d*) Settlers. Sewell and Davidsen added that these patterns of adjustment are related to the sojourner's perceptions of roles and their expectations with respect to returning home.

The Detached Observers were described as having very little involvement or need for involvement with the host country. Their motives for the sojourn are usually technical and clearly defined, and they form friendships with fellow sojourners rather than with host nationals. They do not show severe adjustment problems, either during the sojourn or on returning home.

The Promoters were described as showing low to moderate involvement in the host country. They are committed to their home culture, but become reluctant participants in the foreign culture as a result of their role perceptions; that is, they believe that they are expected to participate in the host culture. However, they undergo some adjustment problems as a consequence of the discrepancies between their home and host-country commitments. They are described as continuously praising their home countries.

The Enthusiastic Participants were described as detaching themselves from their home country and actively participating in the host culture. They are aware of the temporary nature of the sojourn and apparently try to capitalize on their sojourn experiences. Their motives for the sojourn are not exclusively related to their technical goals. They have relatively few adjustment difficulties in the host country; however, their expectations of the return to the home culture are not clearly defined, which may lead to reentry crises.

The Settlers were described as showing little attachment to their home country. They adapt slowly, but, in time, show a genuine and thorough adaptation to the host country.

In India, Useem (1966) classified wives of Americans according to four patterns of behaviors, namely, (*a*) Copers, (*b*) Cautious, (*c*) Supporters, and (*d*) Fumblers. The Copers, in general, have an affirmative, enthusiastic view of life. They emphasize self-direction, initiate action, and have a high degree of self-esteem, although early in their stay in India they have as many problems as the other American wives. If they are not able to solve their problems through action, they put pressure on their husbands to leave the host country.

The Cautious are neither enthusiastic nor indifferent. They weigh alternatives carefully, and slowly work through their problems. If they are not able to cope with their problems, they look for a more congenial environment.

The Supporters are responsive to direction from others. They are supportive of others, although they do not take active leadership. If placed in highly structured situations, they will have adjustment problems until they learn the cultural patterns; if they are in unstructured situations, they will become disorganized by their environment.

The Fumblers are noncoping, incautious, and nonsupportive. They are neither self-directed nor responsive to direction from others. They are a burden to their husbands and to the local community, are chronic complainers, and blame others for their problems. If they are ultimately unable to cope with their problems, their husbands may be asked to resign.

Useem emphasized that these patterns of behavior would not provide an effective basis for the selection of personnel for overseas assignment, since, with the exception of a few Fumblers, for whom fumbling was a life pattern, these behavior patterns were not consistent over periods of time. An older woman who had been a Cautious type during her lifetime in the United States, for instance, might change into a Coper in India. Also a woman could be a Coper during one assignment, but could become a Cautious type during the next assignment, and all could become Fumblers when placed under a great deal of stress. Because Useem found little evidence that patterns of social behavior were long enduring, he found them to be of little value in predicting behavior.

Smith (1956) reported that sojourners from similar cultures tend to develop common patterns of behavior in the host culture. He observed that such common behavioral patterns are especially characteristic of sojourners from certain developing countries when they travel to the United States, because they often perceive a loss of status during their sojourns. For example, foreign scholars who perceive a general superiority of American science and technology may feel a loss of status when studying in the United States (Beals & Humphrey, 1957; Lambert & Bressler, 1956; Scott, 1956). Lambert and Bressler (1955) described the sensitive areas of the foreign student in the United States and the adjustment problems that may result. Smith (1956) reported that Indian students were often appalled by Americans, because of the latter's apparent ignorance and disapproval of the Indian culture, for instance, the caste system and poverty. The Indian students tended to react in a hostile manner and to retaliate against the Americans by referring to major American problems, for instance, the racial problem. In contrast to the Indian students, Japanese students appeared to react to criticism of their home country by employing "enryo," that is, withdrawal from Americans. Whether a student reacted in a defensive manner, either counterattacking as the Indians or withdrawing as the Japanese, seemed to be related to the sojourner's attitudes. Those sojourners who reacted defensively to criticism of their own countries had negative attitudes toward both the United States and their own countries. The relationship between attitude and defensiveness may be a function of the sojourner's negative attitudes toward his home country; that is, in order to maintain his self-esteem he has to deny his negative attitudes and employ a defense mechanism. Thus, the sojourner's negative attitude toward the host country may be a function of his negative attitude toward his own country and of his defensiveness, rather than of his actual sojourning experience.

Bennett, Passin, and McKnight (1958) described the adjustment patterns of Japanese students in the United States, based upon data derived from interviews and a battery of tests. The adjustment patterns of the students were described in relation to three personality types, namely, the Adjuster, the Constrictor, and the Idealist. The Adjuster is able to adapt to both the Japanese and American cultures and is able to assimilate attitudes from both cultures. The Constrictor is conservative in his attitudes, and his behavior is congruent with orthodox Japanese culture. The Idealist is liberal in his attitudes and is open to cultural change. In another study of Japanese sojourners, Nash and Shaw (1963) described three personality types of Japanese immigrants to Cuba, which he termed the Autonomous, Traditional, and Transitional.

Perlmutter (1954) investigated the relationship between the American sojourner's adjustment and xenophilia, that is, the preference for things and persons foreign. Sojourners were administered a scale consisting of sentences that compared foreign and domestic objects. The Xenophiles were defined as those sojourners who chose the foreign in preference to the domestic object, for instance, "Most European girls make better wives than American girls." In spite of the Xenophile's preference for things foreign, he tends to isolate himself in American enclaves while living in the host country. Perlmutter suggested that the Xenophile tends to fantasize about "a better world abroad," regardless of where he may live. It appears likely that the Xenophile is a dissatisfied person who is trying to maintain his self-esteem by holding his environment responsible for his feelings of dissatisfaction. Pool (1965) has presented a more extensive examination of the significance of xenophilia in understanding the sojourner.

Lundstedt (1963) stressed the importance of personality factors in sojourn research. He maintained that a closed mind and ethnocentrism may limit the attainment of effective overseas adjustment, and that the more effective coping mechanisms in adjustment consist of rational attitudes, universalistic tendencies, open-mindedness, and flexibility. A similar viewpoint was taken by Gardner (1962). He suggested that the "universal communicator" will have the least difficulty in adjusting to another country. The universal communicator was described as having a well-integrated personality, a central organization of an extraverted type, a value system that includes the "values of all," a socialization of cultural universals, and a high degree of sensitivity toward others.

The descriptions of the potentially good adjuster, portrayed by Lundstedt and Gardner, appear to be commonly accepted in the literature on sojourning. However, the description is based primarily upon face validity rather than empirical data. In spite of its appeal to common sense, the description may be oversimplified and possibly incorrect. For example, Bronfenbrenner, Harding, and Gallwey (1958) found evidence that general interpersonal sensitivity is associated with passivity and withdrawal; hence, extending their findings to intercultural interactions, the sojourner who has a high degree of sensitivity may be at a disadvantage in adjusting to a foreign culture. Also, Byrnes (1966) described a number of studies about overseas American technicians, and reported that the evidence did not support an "overseas type," that is, the successful overseas technicians did not show any consistent personality patterns that would distinguish them from unsuccessful technicians. Additionally, in a study by Smith (1966) dealing with Peace Corps Volunteers, scores on authoritarianism showed no relationships with either factor loadings on a general competence pattern or with administrative evaluations of the volunteers in the host country.

The conceptualization of the adjustment of the sojourner in terms of typical behavior or social patterns, and personality typologies, characterizations, traits, and constellations is interesting, entertaining, and somewhat anecdotal. However, it becomes apparent, perhaps because of the diversity of these approaches, that little common ground exists for making lucid or even intelligible comparisons or distinctions between them; the particular patterns, typologies, and traits appear to be as varied as the investigators themselves. Also, some authors seem to describe and emphasize particular behaviors, patterns, or modes of adjusting while ignoring the antecedent, background, personality, or social determinants. Other researchers, however, emphasize the personality or background factors but neglect the behavioral consequents. And others apparently fail to distinguish between antecedents and consequents in relation to the patterns, typologies, and traits with the result being a lack of clarity. Most of the investigators have used an impressionistic approach in establishing the social patterns and personality typologies of sojourners without empirically cross-validating whether their impressions are consistent with other sources of data. Thus, the proposed patterns and personality traits seem to be vague descriptions neither clearly defined nor conceptualized in any comprehensive manner. Although it seems unlikely, these descriptions may have some future value in understanding intercultural adjustment, but at the present time they are little more than hypotheses based on limited data and general impressions.

BACKGROUND AND SITUATIONAL FACTORS

A somewhat greater degree of success in predicting intercultural adjustment has been based on background and situational factors. Smith et al. (1963) reported that the religious affiliation of Peace Corps Volunteers was related to adjustment, that is, morale. However, age, type of undergraduate school, and previous teaching experience showed no relationship to morale. Another investigation of Peace Corps Volunteers found that paternal absence during childhood differentiated significantly between successful and unsuccessful volunteers (Suddenfeld, 1967). The proportion of volunteers from fatherless homes was significantly greater among unsuccessful volunteers.

Smith et al. also reported that morale was higher for Peace Corps Volunteers living in or near the major cities than those living in more rural areas within the host country. However, morale was not related to age of African school, subject matter taught, teaching level, or working alone. In another study, Gullahorn and Gullahorn (1963), using a 13-item self-satisfaction questionnaire to measure adjustment, found that American grantees in Europe were better adjusted than those who stayed in the Near or Middle East.

SOCIAL INTERACTION AND ADJUSTMENT

The degree of social interaction or contact between the host national and the sojourner has been found to be related to the latter's adjustment. For example, Sewell and Davidsen (1961) reported a significant relationship between the social interaction of Scandinavian students with Americans and their satisfaction with their sojourns. And Morris (1960) found that the volume, range, and depth of social interactions of foreign students in the United States were significantly related to their satisfaction with their sojourns.

A psychiatric study of American housewives in Columbia indicated that a lack of social interaction may result in adjustment problems (León, 1963). The women patients manifested the typical neurotic symptoms of anxiety, insomnia, heavy drinking, etc., and León suggested that a woman sojourner may have greater difficulties than a man in adjusting to a foreign country because the woman has less opportunity to interact with others. The professional man can derive satisfaction from his work and engage in social interaction in conjunction with his job, whereas the housewife may have few avenues of social contact open to her. Also Gullahorn and Gullahorn (1963) reported that American grantees who had high adjustment scores tended to maintain contact with host institutions, consult on professional matters with colleagues and students in the host country, and help these hosts to apply for grants in the United States.

The earlier sections of this article described the relationship between adjustment and several variables, that is, patterns of behavior, personality, background, and situational factors. In the sections that follow, some of these variables are seen to be related to adjustment insofar as they either facilitate or inhibit social interaction.

Sewell and Davidsen (1956) determined the degree of social interaction that Scandinavian students had with aspects of American culture. The measure of social contact consisted of the students' self-reports of social interaction with respect to American homes, churches, commercial recreation areas, campus activities, and travel within the United States. Background factors found to be related to social interaction included high socioeconomic status in the student's home country, urban background, previous contact with Americans, and a liberal arts as opposed to a physical science college background.

Major (see Footnote 2) stated that the sojourner's geographical background is an important factor influencing social interaction. He found that European in contrast to Asian students, while sojourning in the United States, had more social interactions with Americans. Similarly, American students in Europe reported more social interactions with host nationals than did Americans in Asia.

Sewell and Davidsen reported other factors that influenced the amount of social interaction of foreign students in the United States. Those students interacted with Americans most frequently whose purposes for coming to the United States were social or cultural rather than professional. Also, those students who had relatively less frequent and less severe frustrating intercultural experiences tended to engage more in social interactions with the hosts. In another study, Goldsen, Suchman, and Williams (1956) investigated the social interaction of Americans on a foreign campus, and found that those students who participated most in the campus life revealed a general pattern of friendliness, and were less critical of specific issues and practices of the United States. However, the degree of social interaction was not related to the social, political, or economic interests and attitudes of the students. Kelman (1962) suggested that intercultural interaction may be increased by the sojourner's participation in an ongoing enterprise, maintaining the self-esteem of the host, and by introducing common goals as the result of combining his effort with the hosts on a particular task. Finally, the results of several studies (e.g., Morris, 1960; Selltiz, Christ, Havel, & Cook, 1963) have indicated a positive relationship between one's fluency in the host language and the extent of social inter-interaction with the hosts.

The studies cited in the previous sections indicate that social interaction is a relevant variable in understanding the adjustment of the sojourner. More specifically, it seems that various background and situational factors influence the sojourner's social interaction with his hosts, which in turn influence his adjustment. However, the term social interaction has not been clearly specified by the majority of investigators with the result that the concept remains somewhat vague and elusive. For example, does social interaction refer to talking to host nationals? To the number of people the sojourner talks to? To the amount or type of talking? Does it refer to the situation in which he is merely present with others without actual conversation going on? Does it refer to the number of hosts that the sojourner meets, knows, or is close friends with? Thus, there are many ways in which a sojourner can interact with his hosts, and it would seem necessary to specify more clearly what is meant by social interaction in order to be able to relate dimensions of social interaction to the sojourner's adjustment.

INTERCULTURAL COMMUNICATION

The study of the adjustment of the sojourner should not emphasize the social interaction or contact experience per se, but rather the dimensions of interpersonal processes that occur during these encounters. What is most important, then, is not the mere occurrence of the contact or interaction, but instead, what actually *happens* during the encounter.

The authors believe that the effective interpersonal functioning of the sojourner, that is, his adjustment, is dependent on the development of understanding between himself and his host. The degree that they are able to engender an effective flow or exchange of information will determine the extent that mutual understanding can develop. Thus the sojourner's adjustment is seen to be a function of effective intercultural communication which occurs between sojourner and host (see Ells[3] for an annotated bibliography of articles and books related to intercultural communication).

Many investigators have suggested that differences between cultures may impede effective intercultural communication. Gardner (1962), for example, maintained that because the culture of the United States is extremely different from the cultures of developing countries, the American is rarely able to "perform this psychological feat of putting one's self in the other's shoes with any degree of efficiency [p. 245]." Gardner, using a sociological approach, discussed these differences between cultures and their implications for intercultural communication. He related his ideas to Max Weber's model in which a society moves from a "traditional authority" to a "rational-legal authority." A technologically advanced country would be located at the rational-legal phase, whereas a developing country would be located at the traditional phase. Social interaction between individuals from countries whose social systems are in different phases could result in numerous misunderstandings, with consequent difficulties in social personal adjustment for the sojourner. A somewhat similar sociological approach taken by Riesman, Glazer, and Denny (1955) contends that countries move from "tradition-directed" stages to "inner-directed" stages, and culminate in "other-directed" stages. Zeleznik (1957) related the theory of Riesman et al. to the adjustment of the sojourner.

An interesting comparison between traditional and technologically advanced cultures has been proposed in terms of role differentiation by Foa and Chemers (1967). They suggest that the family is the nuclear system for establishing the various roles within a given society. The differentiation of roles within other systems, such as those pertaining to education or business, is secondary and rooted in the differentiation of family functions. A traditional culture has relatively few secondary systems in comparison with a more technologically advanced culture, which has many. Furthermore, a traditional culture makes many role distinctions within the family, but few distinctions between the family and secondary systems. Conversely, a technologically advanced culture makes relatively fewer role distinctions within the family, but many distinctions between the family and secondary social systems. Thus, for example, in a traditional culture, status in a particular social system may be determined primarily by one's role within the family, whereas, in a technologically advanced culture, status may vary from one social system to another depending on one's role in the particular system. In order to communicate effectively in an intercultural setting, therefore, it would seem particularly important for the sojourner to carefully differentiate among the roles of the host culture. For example, an American, to be able to communicate optimally in the Middle East, must learn to differentiate more within the family and less between the family and other systems. On the other hand, a Middle Easterner sojourning in the United States would have to learn the opposite, namely, to place less emphasis upon the role differentiation within the family, but more emphasis between the family and other social systems.

The difficulty of "putting one's self into the other person's shoes" is aptly described by an Oriental story related by Adams (1962):

> Once upon a time there was a great flood, and involved in the flood were two creatures, a monkey and a fish. Now the monkey, being agile and experienced, was lucky enough to scramble up a tree and escape the raging waters. As he looked down from his safe perch, he saw the poor fish struggling against the swift current. With the very best of intentions, he reached down and lifted the fish from the water. The result was inevitable [p. 179].

The sojourner, especially if he is an educational or technical adviser, may act much like the monkey in the story, and with the most laudable of intentions may make equally disastrous decisions. The subtle and yet very important differences between cultures may not be understood, with the result that what may be perceived by the sojourner as the "most laudable of intentions," may be perceived by the host-country nationals as an insult instead. An awareness of differences between cultures would seem, then, to be crucial for effective intercultural communication. But developing such awareness is very difficult, primarily because of the subtleties of these differences. As Kluckhohn (1949) stated when he compared a culture to a map, "If a map is accurate and you can read it, you won't get lost; if you know a culture, you will know your way around in the life of a society [p. 28]." In a recent study of the Sinasina culture in New Guinea, Turner (1968) stressed that a sojourner should understand the "cultural theme" of the host culture in order to establish adequate intercultural communication.

---

[3] Ells, S. A bibliography on cross-cultural communications. Unpublished manuscript, Putney, Vermont: Experiment in International Living, School for Intertional Training, 1968.

Schuetz (1944) described the role of a stranger in a group, in which the stranger must modify his frame of reference in order to become more aware of the implicit structure and rules of the group. The stranger cannot rely upon vague notions about the norms and structure of the group if he is to become accepted. Likewise, the sojourner in his role as a stranger to a foreign culture needs to become aware of its implicit rules. This is no doubt a more difficult task for him than were he adjusting to a new group within his own culture.

Hall (1959) and Hall and Whyte (1963) discussed the process of intercultural communication and stated that periods of time spent in another culture may bring to awareness aspects of the sojourner's home culture of which he was previously unaware. A sojourner becomes aware of the many subtle cues that he has been employing within his own country, because the cues are no longer effective for communication within the host culture. An awareness of what were formerly unconscious cues for communication might be an important contribution to effective communication within the host culture, because the awareness of a cue may enable a sojourner to modify his patterns of communication so that they are congruent with the cues of the people of the host country. Therefore, if a sojourner can consciously direct his behavior, the likelihood of his becoming better adjusted would be increased. In terms of the downward slopes and low points of the W curve, adjustment would be evidenced by a reduction in the slope and a reduction in the severity of concomitant culture-shock symptoms.

Intercultural communication can be conceived in terms of verbal communication, nonverbal communication, and complex interrelations between the two. The most obvious mode of communication is through the use of language, and its importance to understanding adjustment has been indicated by a number of studies.

Deutsch and Won (1963), for example, reported that language facility of trainees of the Agency for International Development was a factor influencing the degree of their social satisfaction or adjustment. And in a study of Scandinavian students in the United States, Sewell and Davidsen (1956) reported, also, that fluency in the host language was a necessary condition for a satisfactory adjustment. However, Major (see Footnote 2) reported that the English fluency of foreign students in the United States had little bearing on either their social or academic adjustment. Major suggested that the foreign student's confidence in his use of the host language had a greater influence on his adjustment than actual facility in the host language. Gullahorn and Gullahorn (1966) suggested that confidence in the use of language, regardless of ability, may be a function of a more general confidence in one's ability to interact with others. Within certain cultural contexts the sojourner who has confidence in his ability to speak may be encouraged by foreign nationals whenever he uses the host language. However, Gullahorn and Gullahorn noted that not all countries are equally encouraging of the foreigner who attempts to speak in the host language. For example, whereas a sojourner to the United States will be likely to find his attempts to speak English generally encouraged, in France the sojourner may find his hosts relatively intolerant of misuse of the French language. The sojourner who has strong tendencies toward verbal interaction may find such rebuffs considerably frustrating. Other aspects of language have been covered by Bennett et al. (1958), Feldman (1968), and Watson and Lippitt (1955).

Nonlinguistic as opposed to linguistic aspects of culture have been emphasized by Glenn (1967), Loubert,[4] and Wight, Hammons, and Bing[5] in explaining the barriers to communication that often exist between sojourner and host. And Hall (1959) proposed a theory of culture in which verbal communication is only one of many possible modes of communication. He formulated a "map of culture" which can be represented by a 10×10 matrix of cells. While the axes of the matrix consist of "primary message systems" or systems of communication, the cells in the matrix include the totality of activities in any culture. Thus, Hall conceived of culture as a vast system of communication composed of these message systems into which all of the aspects, activities, and elements of a culture may be categorized. Each cell has associated with it a "meaning" or communication value that is generally understood by the majority of members of a given culture. Since only one cell of the matrix is used to represent the process of verbal communication, all of the remaining cells pertain to messages that occur nonverbally and largely out of awareness. Therefore, nonverbal, in comparison with verbal communication, would seem to be of much greater significance within the culture (see Duncan, 1969, for a recent review of nonverbal behaviors).

Hall contends that the greatest confusion in intercultural communication can be traced to a failure to interpret correctly the subtle cues that precede and surround a verbal communication. A person may understand the actual spoken message, but may be totally unaware of the nonverbal messages that are occurring on the unconscious level. This unconscious communication is described as following certain culturally determined and relatively inflexible "rules," and may be discovered through analyses of nonverbal behavioral patterns. Often the unconscious behaviors become known to the sojourner during the period of his relatively greater adjustment difficulties. In the sojourner's home country many actions which were automatic and primarily unconscious are relatively inappropriate within the host culture and become conscious in the process of adjustment (see David, 1971). According to Hall and Whyte (1963) sojourning has an "inoculation" effect. An American, for example, by virtue of his previous travel experience becomes more aware of his own modes of behavior, which were previously unconscious to him in the United States. This new awareness helps, then, to reduce the obstacles that interfere with understanding in subsequent intercultural encounters.

Apparently a variety of differences between cultures may contribute to misunderstandings in the communication process. Cultural differences with respect to time and space are two examples. Important cultural differences and their implications for intercultural communication may be found in the way that people from various countries conceive time. As described by Cleveland et al. (1960), if an American is invited to a seven o'clock dinner, he will probably arrive between 5 and 15 minutes after seven. A Scandinavian will probably arrive exactly at seven, and it is likely that he will bring a gift. A Latin may arrive at nine o'clock, and an Ethiopian may arrive somewhat later. A Javanese may not arrive at all, although he had politely accepted the invitation in order to prevent the host from losing face. Hall (1959) reported that an American who has to wait 45 minutes for a business appointment will likely be insulted, whereas, in some Latin American countries a wait of 45 minutes would be considered unusually prompt.

The nonverbal communicative significance of spatial behavior and its implications for intercultural communication, in particular, have been discussed by Brein[6] and Hall (1959, 1966, 1968). When two persons from different cultures interact, various misunderstandings may result from each interactant's lack of insight into the other's use of space. Misunderstandings can arise from crosscultural differences in territoriality. The middle-class

---

[4] Loubert, J. D. *The trans-cultural research and training institute (TCI)*. (ONR Tech. Rep. No. HSR-RR-67/7-Cs, Contract Nonr-4346-(00), NR-177-238) McLean, Va.: Human Sciences Research, 1967.

[5] Wight, A. R., Hammons, M. A., & Bing, J. *Cross-cultural and community involvement training*. (Peace Corps Tech. Rep., Contract PC-25-1710) Estes Park, Colorado: Center for Research and Education, 1969.

[6] Brein, M. Psychoproxemics. Unpublished manuscript, University of Hawaii, 1970.

American, for example, believes that he has an inherent right to an office, room, or desk of his own, whereas, his English counterpart does not have the same apparent need to establish such territoriality. Also, the American tends to establish territorial rights in public places, for instance, if he is sitting at a public counter, another person is not expected to attempt to dispossess him of his seat; however, the Arab, according to Hall, believes that a public area really is public, and that a person does not have rights to an area or location simply because he occupies it. In such encounters, any number of potential misinterpretations of behavior could occur between Arabs and Americans.

The conversational interaction distances that people maintain between themselves have been noted to vary across cultures. Arabs, Latin Americans, Indonesian, and African groups, among others, have been described as maintaining generally closer interaction distances among themselves than Americans do (Argyle & Dean, 1965; Hall, 1959, 1963, 1966; Hall & Whyte, 1963). Hall suggested that since cultures apparently vary in their norms for interaction distance, people from these cultures structure their spatial experiences differently and may not necessarily attach the same meanings to their space usages. Hall emphasized the implications these differences might have with respect to intercultural communication by providing an example of the misunderstandings that could result from the different cultural meanings attached to the use of space. When a Latin American and a North American converse, both try to maintain their normal conversation distances with

> the result . . . that when they [Latins] move close, we withdraw and back away. As a consequence they think we are distant or cold, withdrawn and unfriendly. We, on the other hand, are constantly accusing them of breathing down our necks, crowding us, and spraying our faces [1959, p. 164].

Although Hall's observations are largely impressionistic, some experimental evidence tends to support his views regarding interaction distances (Brein, see Footnote 6; Forston & Larson, 1968; Sommer, 1967; Watson & Graves, 1966). Other investigators have studied crosscultural differences in spatial behavior by using object arrangement procedures. Based on the methods of Kuethe (1962a, 1962b, 1964), subjects arranged pairs of dolls or paper human cutouts into simulated conversational interactions in any manner they wished onto backgrounds or fields. The distances resulting between these figures have been found to correspond to the actual distances that persons maintain between themselves during conversational encounters. Little (1965, 1968) termed these spatial separations "reproduced interaction distances," and observed that Mediterranean groups (Greeks and Italians), in comparison with Northern European groups (Swedes and Scots), placed dolls significantly closer together during simulated "conversations." Engebretson,[7] in a similar study reported cultural differences between Americans and Japanese. It seems, then, that such object-arrangement tasks may be potentially valuable for studying interaction distances in relation to the processes of intercultural communication.

Triandis, Vassiliou, and Nassiakou (1968), in a cross-cultural study of social perception, suggested that two persons from different cultures interacting with each other could develop misunderstandings and misinterpretations of each other's behaviors as a result of differences in their perceptions of social behaviors. It is possible that they could perceive their shared encounter "in very different terms" with respect to the exchange of reinforcements occurring between them, the degree of intimacy of the encounter, as well as their relative differences in status. For example,

> a person from one culture may provide what he considers to be friendly criticism to a person from another culture only to discover that the other person interprets it as "hatred." Or, a person from culture A behaves in a manner which he considers extremely "positive" toward a person from culture B. However, the individual from culture B perceives the behavior as "neutral," and in turn, the individual from culture A feels that he is "given the cold shoulder" [p. 33].

Triandis et al. also gave a specific example of Greek and American differences in the perception of certain kinds of social behaviors, and discussed the implications of these differences for intercultural interactions between Americans and Greeks.

---

[7] Engebretson, D. E. Cross cultural variations in territoriality: A base-line determination of interactional distance between shared culture dyads. Unpublished doctoral dissertation, University of Hawaii, 1969.

Fayerweather (1959) described how Americans feel complimented when they are told that they are open, direct, and approachable; however, Mexicans are seen to regard openness to be a form of weakness or treachery. The Mexican can give in or humiliate himself, but he should never allow the outside world to penetrate his intimate being. The person who is open is not to be trusted. He is a traitor or a man of doubtful fidelity, one who tells secrets and is incapable of facing dangers as they should be faced. In an article about the first Peace Corps Volunteers to the Philippines, Guthrie (1966) described how the volunteers were urged by some members of the staff to "be perfectly frank" in discussions with the Filipinos. This was a devastating piece of advice. The frankness of the Americans resulted in disrupted interpersonal relationships with the Filipinos because Philippine social interaction depended upon "smooth interpersonal relationships." Szanton (1966) described smooth interpersonal relationships as the handling of potential disagreements by ignoring or denying them, or calling on a third party to negotiate a compromise. Adams (1962) discussed some of the differences between value orientations of Koreans and Americans and also discussed the implications of these differences as they pertain to the role of an American educational adviser.

According to Cleveland et al. (1960), the American sojourner may have difficulty adapting to a culture in which indirection and innuendo are a way of life. For example, in some Asian languages the word for "no" is rarely used or is the same word as "yes"; thus, the Oriental "yes" may mean "maybe" or "no." The Oriental may consider it better to say an untruth than to cause another person to lose face. Cleveland et al. presented a story about an Indonesian middle-class boy who wanted to marry an upper-class girl. The boy's mother went to the house of the girl's mother for tea. A banana was served with the tea, which was a most unusual combination. The women did not discuss the marriage during their visit, but the boy's mother knew that the marriage was unacceptable—bananas do not go with tea. The relevant information had been communicated, but nobody had lost face. However, an American might have picked up the banana and started to peel it while discussing the wedding date.

The preceding examples of intercultural misunderstandings suggest that the adjustment of the sojourner is related to those areas in which differences occur between cultures. Thus, the basis for ineffective communication of the sojourner would seem to be embedded in the differences between the home and host cultures; consequently, further investigations of cross-cultural differences should be able to establish the areas in which intercultural communication may be improved. More specifically, to understand the adjustment difficulties of an individual sojourner it would be necessary to relate specific background factors, personality traits, and situational factors to the communication process. Relating these factors to intercultural communication may then provide an integrating frame of reference by which any number of independent approaches could be compared and understood in relation to the adjustment of the sojourner. In turn, the effective exchange of information would allow for improved intercultural encounter to develop, resulting in improved adjustment for the sojourner. Although effective intercultural communication seems to establish the necessary basis for adjustment, it does not insure that the sojourner will show good adjustment. A sojourner may know how to communicate effectively in the host culture, but he may not be able to either accept or reconcile the differences between cultures. Although it is beyond the scope of this paper, the next step in understanding the adjustment of the sojourner may be related to his acceptance of the host culture once he has developed effective patterns of communication.

REFERENCES

Adams, D. Cultural pitfalls of a foreign educational adviser. *Practical Anthropology*, 1962, 9, 179–184.

Argyle, M., & Dean, J. Eye contact, distance, and affiliation. *Sociometry*, 1965, 28, 289–304.

Beals, R. L., & Humphrey, N. *No frontier to learning: The Mexican student in the United States*. Minneapolis: University of Minnesota Press, 1957.

Bennett, J. W., Passin, H., & McKnight, R. K. *In search of identity: The Japanese overseas scholar in America and Japan*. Minneapolis: University of Minnesota Press, 1958.

Bronfenbrenner, U., Harding, J., & Gallwey, M. The measurement of skill in social perception. In D. C. McClelland, A. L. Baldwin, U. Bronfenbrenner, & F. L. Strodtbeck (Eds.), *Talent and society*. Princeton: Van Nostrand, 1958.

Byrnes, F. C. Role shock: An occupational hazard of American technical assistants abroad. *The Annals*, 1966, **368**, 95–108.

Cleveland, H., Mangone, G. J., & Adams, J. G. *The overseas Americans*. New York: McGraw-Hill, 1960.

David, K. H. Non-verbal communication in an intercultural setting and self-awareness. *Voices: The Art and Science of Psychotherapy*, 1971, in press.

Deutsch, S. E., & Won, G. Y. M. Some factors in the adjustment of foreign nationals in the United States. *Journal of Social Issues*, 1963, **19**(3), 115–122.

Dicken, C. Predicting the success of Peace Corps community development workers. *Journal of Consulting and Clinical Psychology*, 1969, **33**, 597–606.

Duncan, S., Jr. Nonverbal communication. *Psychological Bulletin*, 1969, **72**, 118–137.

Ekroth, L. E. The study of face to face communication between cultures: Present status and directions. *Dissertation Abstracts*, 1968, **28**, 5167A–5168A.

Fayerweather, J. M. *The executive overseas*. Syracuse: Syracuse University, 1959.

Feldman, R. E. Response to compatriot and foreigner who seek assistance. *Journal of Personality and Social Psychology*, 1968, **10**, 202–214.

Fisher, J., Epstein, L. J., & Harris, M. R. Validity of the psychiatric interview. *Archives of General Psychiatry*, 1967, **17**, 744–750.

Foa, U. G., & Chemers, M. M. The significance of role behavior differentiation for cross-cultural interaction training. *International Journal of Psychology*, 1967, **2**, 45–57.

Forston, R. F., & Larson, C. V. The dynamics of space: An experimental study in proxemic behavior among Latin Americans and North Americans. *Journal of Communication*, 1968, **18**, 109–116.

Foster, G. M. *Traditional cultures: And the impact of technological change*. New York: Harper & Row, 1965.

Gardner, G. H. Cross-cultural communication. *Journal of Social Psychology*, 1962, **58**, 241–256.

Glenn, E. S. The delicate art of interpretation. *Trans-action*, 1967, **4**, 30–31.

Goldsen, R. K., Suchman, E. A., & Williams, R. M. Factors associated with the development of cross-cultural social interaction. *Journal of Social Issues*, 1956, **12**(1), 26–32.

Gullahorn, J. E., & Gullahorn, J. T. American students abroad: Professional versus personal development. *The Annals*, 1966, **368**, 43–59.

Gullahorn, J. T., & Gullahorn, J. E. An extension of the U-curve hypothesis. *Journal of Social Issues*, 1963, **19**(3), 33–47.

Guthrie, G. M. Cultural preparation for the Philippines. In R. B. Textor (Ed.), *Cultural frontiers of the Peace Corps*. Cambridge: MIT Press, 1966.

Hall, E. T. *The silent language*. New York: Doubleday, 1959.

Hall, E. T. Proxemics: A system of notation of proxemic behavior. *American Anthropologist*, 1963, **65**, 1003–1026.

Hall, E. T. *The hidden dimension*. Garden City, N. Y.: Doubleday, 1966.

Hall, E. T. Proxemics. *Current Anthropology*, 1968, **9**, 83–108.

Hall, E. T., & Whyte, W. F. Intercultural communication: A guide to men of action. *Practical Anthropology*, 1963, **10**, 216–299.

Jacobson, E. H. Sojourn research: A definition of the field. *Journal of Social Issues*, 1963, **19**(3), 123–129.

Kelman, H. C. Changing attitudes through international activities. *Journal of Social Issues*, 1962, **18**(1), 68–87.

Kelman, H. C. (Ed.) *International behavior*. New York: Holt, Rinehart & Winston, 1965.

Kluckhohn, C. *Mirror for man*. New York: McGraw-Hill, 1949.

Kuethe, J. L. Social schemas. *Journal of Abnormal and Social Psychology*, 1962, **64**, 31–38. (a)

Kuethe, J. L. Social schemas and the reconstruction of social object displays from memory. *Journal of Abnormal and Social Psychology*, 1962, **65**, 71–74. (b)

Kuethe, J. L. Pervasive influence of social schemata. *Journal of Abnormal and Social Psychology*, 1964, **68**, 248–254.

Lambert, R. D., & Bressler, M. The sensitive area complex: A contribution to the theory of guided culture contact. *American Journal of Sociology*, 1955, **60**, 583–592.

Lambert, R. D., & Bressler, M. *Indian students on an American campus*. Minneapolis: University of Minnesota Press, 1956.

León, C. A. Problemas adaptivos de un grupo de amas de casa Norteamericanas en Latinoamerica. *Acta Psiquiatrica y Psicologica de Argentina*, 1963, **9**, 114–121.

Little, K. B. Personal space. *Journal of Experimental Social Psychology*, 1965, **1**, 237–247.

Little, K. B. Cultural variations in social schemata. *Journal of Personality and Social Psychology*, 1968, **10**, 1–7.

Lundstedt, S. An introduction to some evolving problems in crosscultural research. *Journal of Social Issues*, 1963, **19**(3), 1–9.

Lysgaard, S. Adjustment in a foreign society: Norwegian Fulbright grantees visiting the United States. *International Social Science Bulletin*, 1955, **7**, 45–51.

Mischel, W. Predicting the success of Peace Corps volunteers in Nigeria. *Journal of Personality and Social Psychology*, 1965, **1**, 510–517.

Morris, R. T. *The two-way mirror: National status in foreign students' adjustment*. Minneapolis: University of Minnesota Press, 1960.

Nash, D., & Shaw, L. C. Personality and adaptation in an overseas enclave. *Human Organization*, 1963, **21**, 252–263.

Oberg, K. Cultural shock: Adjustment to new cultural environments. *Practical Anthropology*, 1960, **7**, 177–182.

Patterson, W. D. *The big picture*. New York: American Society of Travel Agents, 1969.

Pearson, D. The Peace Corps volunteer returns—problems of adjustment. *Saturday Review*, 1964, **47**, 54–56.

Perlmutter, H. V. Some characteristics of the xenophilic personality. *Journal of Psychology*, 1954, **38**, 291–300.

Pool, I., de S. Effects of cross-national contact on national and international images. In H. C. Kelman (Ed.), *International behavior*. New York: Holt, Rinehart & Winston, 1965.

Riesman, D., Glazer, N., & Denny, R. *The lonely crowd*. New York: Doubleday, 1955.

Schuetz, A. The stranger: An essay in social psychology. *American Journal of Sociology*, 1944, **49**, 499–507.

Scott, F. D. *The American experience of Swedish students: Retrospect and aftermath*. Minneapolis: University of Minnesota Press, 1956.

Selltiz, C., Christ, J. R., Havel, J., & Cook, S. W. *Attitudes and social relations of foreign students in the United States*. Minneapolis: University of Minnesota Press, 1963.

Sewell, W. H., & Davidsen, O. M. The adjustment of Scandinavian students. *Journal of Social Issues*, 1956, **12**(1), 9–19.

Sewell, W. H., & Davidsen, O. M. *Scandinavian students on an American campus*. Minneapolis: University of Minnesota Press, 1961.

Shelton, A. J. The "Miss Ophelia" syndrome in African field research. *Practical Anthropology*, 1964, **11**, 259–265.

Smalley, W. A. Culture shock, language shock, and the shock of self-discovery. *Practical Anthropology*, 1963, **10**, 49–56.

Smith, M. B. Some features of foreign-student adjustments. *Journal of Higher Education*, 1955, **26**, 231–241.

Smith, M. B. A perspective for further research on cross-cultural education. *Journal of Social Issues*, 1956, **12**(1), 56–67.

Smith, M. B. Explorations in competence: A study of Peace Corps teachers in Ghana. *American Psychologist*, 1966, **21**, 555–566.

Smith, M. B., Fawcett, J. T., Ezekial, R., & Roth, S. A factorial study of morale among Peace Corps teachers in Ghana. *Journal of Social Issues*, 1963, **19**(3), 10–32.

Sommer, R. Small group ecology. *Psychological Bulletin*, 1967, **67**, 145–152.

Stolley, R. B. Re-entry crisis: Return of Peace Corps volunteers to U. S. *Life*, 1965, March, 98–100+.

Sudenfeld, P. Paternal absence and overseas success of Peace Corps volunteers. *Journal of Consulting Psychology*, 1967, **31**, 424–425.

Szanton, D. L. Cultural confrontation in the Philippines. In R. B. Textor (Ed.), *Cultural frontiers of the Peach Corps*. Cambridge: MIT Press, 1966.

Thomson, C. P., & English, J. T. Premature return of Peace Corps volunteers. *Public Health Reports*, 1964, **79**, 1065–1073.

Triandis, H. C., Vassiliou, V., & Nassiakou, M. Three crosscultural studies of subjective culture. *Journal of Personality and Social Psychology*, 1968, **8** (4, Part 2).

Turner, C. V. The Sinsana "Big Man" complex: A central culture theme. *Practical Anthropology*, 1968, **15**, 16–23.

United States Bureau of the Census. *Statistical abstract of the United States*. (90th ed.) Washington, D. C.: United States Government Printing Office, 1969.

Useem, R. H. The American family in India. *The Annals*, 1966, **368**, 132–145.

Watson, O. M., & Graves, T. D. Quantitative research in proxemic behavior. *American Anthropologist*, 1966, **68**, 971–985.

Watson, J., & Lippitt, R. *Learning across cultures: A study of Germans visiting America*. Ann Arbor: University of Michigan Press, 1955.

Zeleznik, C. Inner and other directedness of the American abroad. *Phylon Quarterly*, 1957, **19**, 258–267.

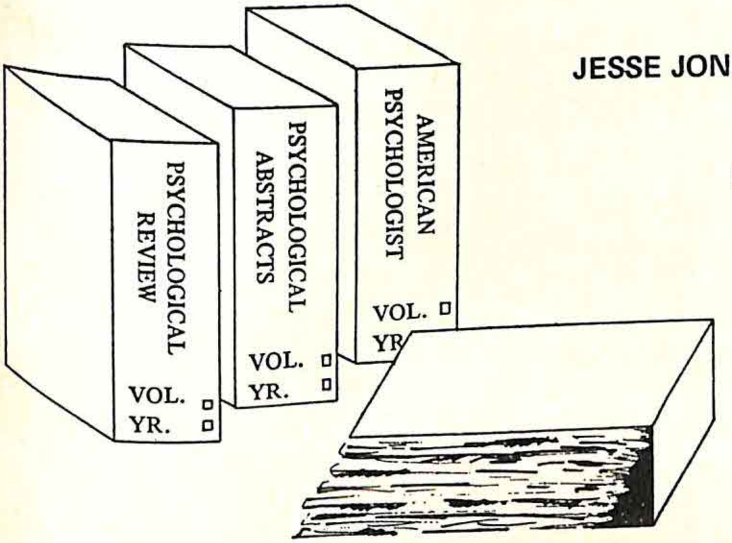
(Received March 16, 1970)

Manuscripts Accepted for Publication in the *Psychological Bulletin*

Reconsideration of a Measure of Power Motivation: Joseph Veroff* and Joanne B. Veroff: Department of Psychology, University of Michigan, Ann Arbor, Michigan 48104.

---

* Asterisk indicates author for whom the address is supplied.



JESSE JONES VOLUME FILES

Send orders to:
JESSE JONES CORP.
P.O. Box 5120
Philadelphia, Pa. 19141

Only $3.50 ea.
3 for $10.00
6 for $19.00 P.P.D.

Now, the journals of the American Psychological Association can become an attractive permanent part of your professional library. These famous Jesse Jones volume files, especially designed to keep your copies orderly, readily accessible for future references — guard against soiling, tearing, wear or misplacement of copies.

This durable file will support 150 lbs. Looks and feels like leather and is washable. The 23-carat gold lettering makes it a fit companion for the most costly binding.

NOW AVAILABLE

A SOURCE BOOK FOR THE FIRST COURSE IN PSYCHOLOGY

A source book for teachers of general psychology, containing reviews of introductory texts, books of readings, and lab manuals; descriptions of psychological journals; a catalogue of audiovisual materials; a listing of psychological equipment manufacturers; bibliographies of collateral readings for students and reference works for teachers; names and addresses of national organizations; a discussion of teaching methods; and suggestions of ways to organize a first course in psychology. The book is one of several products of a project held at Oberlin College in the summer of 1970. Although the title of the book — *PROGRAM ON THE TEACHING OF PSYCHOLOGY IN THE SECONDARY SCHOOL:* Final Report indicates an applicability to a particular academic level, the book is useful in teaching general psychology regardless of the level.

Pages: 128

Price: $2.00

SOURCE BOOK ORDER FORM

Name ....

Address ....

....

....ZIP CODE....................

Send this form to:

American Psychological Association
Order Department
1200 Seventeenth Street, N.W.
Washington, D. C. 20036

Orders must be prepaid. Single copy $2.00. Quantity discounts: 3-10 copies 10%, 11-Over copies 20%. Make checks payable to the American Psychological Association.

VOL. 76, NO. 4 OCTOBER 1971

Psychological Bulletin

REVIEW AND THEORETICAL ANALYSIS OF THE EFFECTS OF GROUP PROCESSES UPON INDIVIDUAL AND GROUP DECISIONS INVOLVING RISK[1]

AMIRAM VINOKUR[2]

University of Michigan

A theoretical organization of studies comparing individual and group decisions involving risk is provided, emphasizing what has been called the "risky-shift phenomenon." Proposed explanations are categorized according to four underlying processes: affective, cognitive, interactive, and statistical. A review of these explanations suggests that only three specific explanations of the shift phenomenon are viable: two complementary cognitive hypotheses (value—information relevant to the task; and rationality—subjective expected utility) and one interactive hypothesis (extremity and influence). An interpretation, consistent with all three hypotheses and akin to the partially shared information model of group productivity and the group problem-solving model, is offered. The theoretical significance of studies concerning individual and group risk taking is discussed, and some implications for further research are drawn. The use of behavioral decision theory—in particular, the subjective expected utility model—for complex, choice-dilemma situations is advocated as the most promising approach for further research in this area.

The bulk of work on the risky-shift phenomenon is characterized by the eclectic nature of the ad hoc hypotheses proposed to explain it. A consistent and systematic approach based on a broad theoretical framework has not been attempted (for reviews see Clark, 1971; Kogan & Wallach, 1967a). Consequently, even if some explanation of the shift phenomenon has been provided, theoretically meaningful new questions about group processes have not been generated. The present review provides a theoretical framework that organizes the results of past research and raises new theoretical and empirical issues. These concern not only risky and conservative shifts but also the underlying group processes.

A comprehensive theoretical approach should begin with a close examination of the data concerning the phenomenon in question. In particular, two characteristics of the risky-shift phenomenon have not received adequate attention. First, while the generality of the risky shift has been documented, its limits are rarely acknowledged. Among the 12 hypothetical choice dilemmas in the questionnaire used in the original risky-shift study by Stoner[3], there are two items (5 and 12) that consistently yield group shifts toward greater caution (conservatism), though the magnitudes of these shifts are at best moderate (i.e., a third to a half of the mean risky shift). Other hypothetical choice-dilemma items that yield cautious shifts have been constructed by various researchers (see Burns[4]; Myers[5]; Nordhøy[6];

---

[1] Preparation of this manuscript was supported by a predoctoral Ford Foundation Fellowship administered by the Rackham Graduate School of the University of Michigan and in part by a grant from the National Institute of Mental Health Grant MH 16950-01. The author is grateful to Eugene Burnstein, Lawrence Mohr, Icek Ajzen, and Phillip Shaver for valuable critical comments on the manuscript.

[2] Requests for reprints should be addressed to Amiram Vinokur, Research Center for Group Dynamics, Institute for Social Research, University of Michigan, Ann Arbor, Michigan, 48106.

[3] Stoner, J. A. F. A comparison of individual and group decisions involving risk. Unpublished master's thesis, School of Industrial Management, Massachusetts Institute of Technology, 1961.

[4] Burns, J. F. An extremity-variance analysis of group decisions involving risk. Unpublished doctoral dissertation, Massachusetts Institute of Technology, 1967.

[5] Myers, D. G. Enhancement of initial risk taking

© 1971 by the American Psychological Association, Inc.

Rabow, Fowler, Bradford, Hofeller, & Shibuya, 1966; Stoner, 1968; Vidmar & Burdeny, 1969).

Another important limitation on the generality of risky-shift phenomenon is its failure to show up in studies of betting behavior (Collins & Guetzkow, 1964; Hunt & Rowe[7]; Lonergan & McClintock, 1961). Similarly, the risky shift has not appeared consistently in Estes-type gambling situations where subjects "bet" which of two lights would appear. Two such studies found cautious shifts (Smith & Radinsky[8]; Zajonc, Wolosin, Wolosin, & Sherman, 1968); a third obtained both cautious and risky shifts under different conditions (Zajonc, Wolosin, Wolosin, & Sherman, 1969). Two recent studies have produced the risky shift in betting, but their results are not unequivocal. In one study (Pruitt & Teger, 1969), the risk-taking scale was biased toward riskiness, as subjects could not choose alternatives with odds of success greater than 1 in 2 except for one alternative with sure success, that is, odds of 1 in 1. The second study (Blank, 1968), unlike any other demonstrating the risky shift, included choices whose expected values increased with their riskiness.

Second, it is also important to note that in studies demonstrating the risky shift, risk is defined only in terms of the odds of success (i.e., probability) of an alternative course of action. The smaller the odds of success, the riskier the decision is considered to be. Hence, the risky-shift phenomenon refers to an apparent change in preference for smaller odds of success. This, however, is not and should not be the only viable definition of risk-taking behavior. In some studies, risk taking has also been defined as choice of an alternative with greater variance of probabilities and values (Coombs & Pruitt, 1960, 1961) or as choice of an alternative with greater uncertainty about these probabilities and values (Marquis & Reitz, 1969). For reviews concerning the definition and assessment of risk-taking behavior, see Slovic (1964) and Coombs and Huang.[9]

If we are to develop an adequate theory of individual and group decisions involving uncertainty, it should include an explanation of the decision-making process and should account for both risky and conservative shifts. It should also go beyond the narrow definition of risk so far employed, and be applicable to a greater variety of tasks. The rudiments of such a theory emerge from the following organization and review of the literature.

The various explanations that have so far been offered to account for the risky shift can be organized according to the processes assumed to underlie the effect. Four types of hypothetical processes can be distinguished: affective, cognitive, interactive, and statistical.

AFFECTIVE HYPOTHESES

Making a decision in a group can produce positive or negative consequences for the individual. If he takes these consequences into account, and they influence his decision, an affective process is said to be operative. In other words, if what the individual may lose or gain by virtue of the fact that he is making his decision in a particular context (e.g., under group pressure, in the presence of others, under pressure to reach a quick decision) influences his decision, the process of influence is an affective one. Borrowing from the terminology of behavioral decision theory (Edwards Lindman, & Phillips, 1965), it can be stated that affective processes involve a shift from one payoff-probability decision matrix to another, where the consequences of making the decision in a group enter into consideration.

It has been argued (Goodman)[10] that a change

---

tendencies in social situations. Unpublished doctoral dissertation, University of Iowa, 1967.

[6] Nordhøy, F. Group interaction in decision-making under risk. Unpublished master's thesis, School of Industrial Management, Massachusetts Institute of Technology, 1962.

[7] Hunt, E. B., and Rowe, R. R. Group and individual economic decision making in risk conditions. In D. W. Taylor (Ed.), *Experiments on decision making and other studies.* (Tech. Rep. No. 6, AD 253952) Arlington, Va.: Armed Services Technical Information Agency, 1960.

[8] Smith, G. F., and Radinsky, T. L. The effect of diffusion and concentration of responsibility on the risky shift phenomenon in a two-choice situation. Paper presented at the Spring meeting of the Iowa Academy of Science, Iowa City, April 1969.

[9] Coombs, C. H., and Huang, L. C. *A portfolio theory of risk preference.* (Rep. No. 68-5) Ann Arbor: Michigan Mathematical Psychology Program, 1968.

[10] Goodman, B. Group decision making in situations of risk. Unpublished manuscript, Univerity of Michigan, 1967.

in a decision about risk level resulting from an affective process is temporary, and therefore should disappear when the individual leaves the group. The permanency of a change is not, however, sufficient evidence that affective processes were not operative. Once a change has occurred, intervening factors may work to stabilize it. Commitment to the decision is just one case in point. Similarly, the interaction among group members cannot be taken as either a necessary or a sufficient condition that an affective process was operative. The individual may be affected by anticipated consequences that are not mediated by the interaction as much as he may recognize those consequences but not let them influence his decision, or he may fail to be aware of them altogether.

The hypotheses that fall in this category are diffusion of responsibility, familiarization, and risk as a value (i.e., information about other's choices).

Diffusion of Responsibility

Studies relevant to this hypothesis have been reviewed by Clark (1971). Though the evidence weighs against this hypothesis, it is interesting to note that the supporting evidence that does exist is equally compatible with an opposite hypothesis. It can be argued that subjects are willing to take greater risks in groups because of the potential prestige that will befall them if their choice leads to success. In other words, the emphasis here is on rewards for success rather than blame for failure. This might be called the "assuming of responsibility hypothesis." According to this hypothesis, individuals in groups will tend to take greater risks in order (*a*) to demonstrate their willingness to assume responsibility and thereby achieve a position of leadership, and (*b*) to take credit for potential success. Intuitively the two hypotheses are equally plausible, and from the research conducted so far it is impossible to distinguish them.

Familiarization

According to this hypothesis, when a subject is pressed to make a decision he will take a cautious position. Upon further familiarization, his initial cautiousness will evaporate, allowing him to endorse a decision involving greater risk. Thus it is assumed that the process responsible for the risky shift occurs before the individual is even exposed to the group or engaged in familiarization with the problem, that is, when he takes an initially rather cautious decision. Since the subject's decision is influenced by the circumstances under which the decision is being made, the underlying process is affective, although it also has some cognitive overtone. Since the bulk of the evidence is against the hypothesis, it suggests that the risky shift is not a pseudogroup effect (Clark, 1971), but is actually mediated by some kind of group process.

Risk-as-a-Value Hypothesis: Information about Others' Choices

This hypothesis has undergone several modifications since first being advanced by Brown in 1965. In its general form, the hypothesis maintains that riskiness is valued in American culture. Hence, risk taking is socially desirable, and those who take risks conform to a cultural norm emanating from the cultural value of riskiness. Brown (1965) maintained that the consequences of the cultural value of risk in a group situation are twofold. First, the cultural value of risk biases the discussion and leads group members to raise more information and generate more arguments favoring the risky alternative. This biased flow of information was first noticed and reported by Nordhøy (see Footnote 6). Second, the cultural value of risk leads subjects to believe that they are somewhat riskier than others and thus closer than the average person to their cultural ideal. This was initially observed and reported by Hinds.[11] Each of these two consequences serves as a basis for a different version of the value hypothesis. Although usually ignored in discussions of the value hypothesis (e.g., Clark, 1971), the distinction between these two versions is of considerable importance since the assumed underlying processes are quite different.

The first version maintains that the biased flow of information and arguments favoring the risky alternative influence subjects to take

[11] Hinds, W. C. Individual and group decisions in gambling situations. Unpublished master's thesis, School of Industrial Management, Massachusetts Institute of Technology, 1962.

greater risks. They are assumed to be convinced on the basis of the arguments that a more risky decision is what is called for in those situations. The second version, often referred to as information exchange or information about others' choices, argues that during the discussion, subjects who believe themselves to be more risky than others are exposed to the choices of others. They soon learn that relative to the others, they are in fact less risky than they thought they were. In view of this discovery, to measure up to what the cultural value of risk demands, they change their prior decision and take a greater risk. Whereas the first hypothesis implies the operation of a cognitive process—subjects are convinced by virtue of the arguments—the second hypothesis implies the operation of an affective process—subjects are influenced by the circumstances of the situation under which the decision is being made, where social comparison process is the specific mechanism (Festinger, 1954). Brown committed himself to the affective hypothesis (information about others' choices), believing that information about others' choices is the primary cause of the risky shift.

> The content of the discussion, the arguments pro and con are of no importance by this theory. It is the information about people's answers that makes individuals move toward greater risk after group discussion [Brown, 1965 p. 702].

In order to account for the conservative shifts found on Items 5 and 12 of the choice-dilemmas questionnaire and on other items written by Nordhøy, Brown revised the hypothesis. Thus it was hypothesized that any given problem involving risk elicits widely held cultural values that favor either the risky or the cautious alternative. On problems for which the cultural values favor a risky decision, the individuals will tend to be rather risky from the beginning and consider themselves to be more risky than most others. In the discussion, they will find out that others are in fact riskier than they are, and this will lead them to change and endorse greater risk. The converse will be true for problems for which the widely held cultural values favor the cautious decision.

Most of the evidence that is generally cited in support of this version of the hypothesis is equivocal and inconclusive (Clark, 1971). The first evidence that is often alleged to support this hypothesis, as well as the other version of the hypothesis, is the product-moment correlation between the initial overall mean risk level, and the mean shift toward risk across the 12 items of the choice-dilemmas questionnaire that was used by Stoner (see Footnote 3) in the original risky-shift study. High correlations were reported by Wallach, Kogan, and Bem (1962), Teger and Pruitt (1967) and Pruitt and Teger[12] to be $r = .83$, $.78$, $.81$, respectively. Another correlation was reported by Stoner (1968) to be $r = .89$ ($p < .001$). Stoner used a different set of 12 items, including 4 items that yielded cautious shifts. These correlations are usually seen as support for the hypothesis, since they are supposed to indicate that on items yielding large shifts toward risk the initial risk-level mean is on the risky side of the continuum (i.e., low odds), and on items yielding conservative shifts the mean is on the conservative side of the continuum (i.e., high odds).

Unfortunately, what the correlations are supposed to indicate is not quite what one finds upon closer examination of the data. For the items that yielded cautious shifts the initial mean risk level is indeed on the conservative side of the continuum. For the two conservative items of the choice-dilemmas questionnaire (5, 12; i.e., those that yielded cautious shifts), the average probability is .75 (see Pruitt & Teger, Footnote 12, for pooled results based on six earlier studies producing the risky shift). For the four conservative items used by Stoner, the average is .76 (Stoner, 1968). But, for the 10 risky items of the choice-dilemmas questionnaire (i.e., those that yielded risky shifts) the average probability is .54 (Pruitt & Teger, see Footnote 12). Even if only the four most risky items are taken (i.e., those that produced the greatest shift toward risk), this average is still .49. Only in Stoner's (1968) study is the mean of the initial risk-level choices on the risky items somewhat on the risky side of the continuum, with an average of .42. These results provide little support for the hypothesis. They make sense only with respect to the conservative shifts, but certainly not for the more common risky shift on the choice-dilemmas questionnaire.

---

[12] Pruitt, D. G., and Teger, A. I. Is there a shift toward risk in group discussion? If so, is it a group phenomenon? If so, what causes it? Paper presented at the annual meeting of the American Psychological Association, Washington, D. C., September 1967.

The second kind of evidence in support of this hypothesis illuminates the social comparison mechanism. Before being exposed to the choices of others during the discussion, subjects considered themselves to be somewhat more risky than most others on items yielding risky shifts, and more conservative on items yielding cautious shifts (Hinds, see Footnote 11; Levinger & Schneider, 1969; Stoner, 1968; Wallach & Wing, 1968; Willems, 1969). During the group discussion, the discovery of their inaccurate assessment of how they stand with respect to others and the elicited cultural norm influences them to take greater risk on risky items and smaller risk on cautious ones. There is, however, an entirely different explanation for these results that is considered later.

The third kind of evidence advanced to support the hypothesis is that a high risk taker be viewed more favorably than his conservative counterpart on items that yielded risky shifts; and also that the conservative risk taker be viewed more favorably than his risky counterpart on items that yielded cautious shifts. Madaras and Bem (1968) compared subjects' ratings of hypothetical risk accepters and risk rejecters on semantic differential scales. These hypothetical figures were faced with dilemmas from the choice-dilemmas questionnaire and other sources, which included five items that have been shown to yield conservative shifts. They found that for the risky items the risk accepter was indeed viewed more favorably than the risk rejecter. The risk accepter was rated as more strong, active, successful, fast, hard, and masculine, and somewhat more good and sociable than the risk rejecter. But on the cautious items, the risk rejecter was not rated more favorably than the risk accepter. In fact, the risk accepter was still rated more favorably than the risk rejecter, although the difference was much smaller on most dimensions and nonexistent on some. Similar evidence was produced by Levinger and Schneider (1969). They found that subjects reported they admired choices riskier than their own, both on the 10 risky items of the choice-dilemmas questionnaire (statistically significant) and on the two conservative ones (not statistically significant). These results provide some direct evidence for Brown's risk-as-a-value hypothesis, that is, that on risky items, risk taking is socially desirable. However, they only *suggest* that the desirability of being perceived as a risk taker by others, or even by oneself, is what leads subjects to take greater risks in a group situation. These results failed to support the part of the hypothesis designed to account for conservative shifts, namely, the contention that on the conservative items, caution is prescribed by the cultural values and thus socially desirable.

The results cited so far are in part compatible with other hypotheses. Therefore, the crucial evidence in support of Brown's hypothesis should be related to the mechanism underlying the shifts. That is, it must be shown that information about others' choices is a sufficient condition for shifts to occur. In a study conducted by Blank (1968), subjects made choices among gambling bets on tossing dice alone and in the presence of others also working on the same task. In the latter condition, subjects made their choices in turn but did not discuss them. The results supported the hypothesis showing greater preference for low-probability (i.e., risky) bets when choices of others were known. In this study, however, the expected value of the low-probability bets was about three to four times greater than that of the high-probability bets, thus casting doubt on the comparability of its results to those of other studies in which expected values were usually constant.

There is accumulating evidence indicating that information about others' choices is not a sufficient condition for shifts to occur; at least not for shifts of the same magnitude as the risky shift after group discussion on the choice-dilemmas questionnaire. Wallach and Kogan (1965) asked subjects to reveal their choices to each other on cards, change them if necessary, and reach a unanimous decision on the choice-dilemmas questionnaire. Subjects were not allowed to discuss the problems. No shift toward risk was found. Though shifts toward greater risk were produced in another study using the same procedure—with the exception of deemphasizing the importance of reaching a unanimous decision—their magnitude was less than half of that found in the condition of discussion to consensus (Teger & Pruitt, 1967).

In another study, subjects made initial decisions on all 12 items of the choice-dilemmas questionnaire and later discussed only half of them, thereby being forced to discover that in general, they were relatively less risky than they had thought. A risky shift was found only on the items that had been discussed; there were no shifts on the other items. These results led the authors of the study to conclude that "the crucial information transmitted in group discussion is not the risk level of others but information relevant to the specific items discussed [Madaras & Bem, 1968, p. 360]." A similar experimental design was used in a study by Pruitt and Teger (1969). In that study, the risk-taking task involved choices among bets. Again, a risky shift was produced only on items that had been discussed. The authors drew a conclusion similar to the former, saying that

> some sort of *communication* among group members about the issue of risk taking is necessary for a risky-shift to occur. This communication may even take the abbreviated form of exchanging information about preferences (Teger & Pruitt, 1967), though the effect is at its strongest when a full group discussion takes place [Pruitt & Teger, 1969, p. 124].

In conclusion, in spite of some indirect supportive evidence, Brown's hypothesis that information about others' choices is sufficient to produce shifts must be rejected. The evidence suggests that the crucial factor responsible for both risky and conservative shifts is something in the flow of information relevant to the issues being decided upon, and not the mere knowledge of others' choices.

COGNITIVE HYPOTHESES

The cognitive hypotheses suggest that the overt change of decision (i.e., shifts toward greater risk or greater conservatism) is the result of a covert cognitive opinion change. These hypotheses are concerned with alterations in decisions that result from changes in the cognitive view of the *situation being evaluated* rather than alterations resulting from *direct influence of the circumstances under which the decision is being made* (e.g., group discussion, presence of others).

Borrowing again from the terminology of behavioral decision theory (Edwards et al., 1965), cognitive hypotheses are concerned with changes *within* the payoff-probability decision matrix that the individual initially considers relevant for the given issue. It is still the same decision matrix that the individual is weighing in the sense of containing the same factors: the relative desirabilities (utilities) of the potential outcomes and the likelihood (probability) that each course of action would lead to those outcomes. Thus, his decision may change as a result of a change in his judgment of the relative utilities of the outcomes and/or their probabilities.

Risk-as-a-Value Hypothesis: Information Relevant to the Task

This hypothesis was considered in the previous section. Recall that the evidence for cautious shifts required that the hypothesis be reformulated. To wit, any given problem elicits widely held cultural values that favor either the risky or the cautious alternative, and thus information flow will be biased in that direction. However, the crux of the matter is that shifts are produced because subjects change their view of the situation by virtue of the information and arguments being generated in the discussion, rather than by mere exposure to the choices of others. In other words, contrary to Brown's contention, the content of the discussion and the arguments "pro" and "con" do constitute the primary factor responsible for the shifts.

Because most of the research on risk as a value was geared to test the information-about-others-choices hypothesis, the information-relevant-to-the-task hypothesis gained only indirect support. This indirect support stems from findings leading to the rejection of the information-about-others-choices hypothesis. There is evidence that some sort of communication in the form of information relevant to the task is necessary for the full strength shift to occur (Madaras & Bem, 1968; Pruitt & Teger, 1969). Also compatible with the information-relevant-to-the-task hypothesis is the evidence that subjects' decisions are strongly affected by changes of information provided in the task, as well as by arguments brought up during the discussion. In one study by Rabow et al. (1966), changes introduced into the problems resulted in changes in the magnitude and dirctions of the shifts. Nordhøy, who originally noticed the biased flow of arguments in favor of riskiness, introduced two small changes on Item 12 of the choice-dilemmas questionnaire that resulted in reversal from a cautious to a risky shift (cited by Brown, 1965). This item involves an engaged couple who sought the advice of a marriage counselor about whether they should get married. A frequent argument in discussions of this item seemed to be persuasive: "If they sought a counselor's advice they are not really in love; therefore, they should not risk marriage unless the odds of successful married life in the future are high." When additional facts were added—the couple were in love and met the counselor incidentally at a party—a shift toward greater risk was observed. In discussions of Item 7 of the choice-dilemmas questionnaire, where the magnitude of the risky shift is greatest, the present author has noticed that a particular persuasive argument is often brought up. This item involves a low-ranked chess player playing against a top-favored player. The question arises whether he should attempt a risky maneuver that might bring him a quick victory. Groups were willing to make riskier decisions whenever someone suggested that the player attempt the risky maneuver, even if the odds of success are low, by arguing that "otherwise he would lose anyway, since as a low-ranked player he is obviously not as good as the top-favored player."

Another line of evidence favoring the hypothesis has already been mentioned. Subjects exhibit the risky shift after listening to real discussions of the choice-dilemmas questionnaire individually (Kogan & Wallach, 1967b; Lamm, 1967); the riskier group decisions are accepted by the subjects as their private decisions (Lamm, 1967; Wallach & Kogan, 1965), and this acceptance endures for at least 6 weeks (Johnston[13]; Wallach et al., 1962) indicating the permanence of the change that led to the shift.

In summary, although all the evidence is at least consistent with the information-relevant-to-the-task hypothesis, the latter has not been subjected to direct test. For a rigorous direct test, it would be necessary to point out the particular information that produces the shifts.

Rationality: Expected Value and Subjective Expected Utility Hypotheses

The most elaborated formulation of the rationality hypothesis was offered by Kogan and Wallach (1967a), though it was initially discussed by Brown (1965). The hypothesis is based on behavioral decision theory (for reviews of behavioral decision theory, see Becker & McClintock, 1967; Edwards, 1961; Edwards et al., 1965). Briefly, behavioral decision theory is concerned with choices among alternative courses of action. The variables that are dealt with in the theory are the values and utilities (personal subjective value) of outcomes resulting from alternative actions, and the probabilities and subjective probabilities that each course of action will lead to a certain outcome. One useful strategy is to choose an alternative that maximizes average gain or minimizes average loss. The average gain or expected value (EV) of a given alternative action is a function of the probabilities (p_i) that certain outcomes will follow the particular action, multiplied by the respective values (v_i) attached to these outcomes, as follows:

$$EV = \sum_{i=1}^{n} p_i v_i, \quad i = 1, 2, \ldots, n \text{ outcomes.}$$

Similarly, when *subjective* average gain is taken to be the criterion for choosing an alternative course of action, it is determined by the multiplication of the subjective probabilities (sp_i) and the respective utilities (u_i) as follows:

$$\text{SEU} = \sum_{i=1}^{n} sp_i u_i,$$

where SEU is subjective expected utility.

Basic to the rationality (expected value) hypothesis regarding the risky shift is the equation of a rational decision with the choice of an alternative that maximizes expected value or subjective expected utility, and the assumption that risk taking represents a more rational maximizing approach in the choice dilemmas that yielded risky shifts, and vice versa for those that yielded conservative shifts. Once this assumption is made, the hypothesis can be stated as follows: A group is in a better position than an individual to unearth information that will aid in the accurate assessment of the probabilities and utilities associated with the choice alternatives. This will result in shifts in the direction required to maximize subjective expected utility. It remains to be shown that group decisions reflect better maximization of expected value or subjective expected utility than individual decisions. (Note that increased maximization of expected value or subjective expected utility implies better utilization of the information already available to individual subjects before the group discussion; it does not require the introduction of new information during the discussion.) On the basis of two studies, Kogan and Wallach (1967a) argued that this is not the case. In one study, expected value was increased for the more conservative alternatives (Bem, Wallach, & Kogan, 1965). In the other study, expected value was the same for all alternatives (Wallach, Kogan, & Bem, 1964; see also Clausen[14]). In both cases, a risky shift was produced. In a more recent study by Pruitt and Teger (1969), the risky shift was also produced in a gambling task involving alternatives with the same expected values. Consequently, the rationality hypothesis must be rejected if rationality is defined as maximization of expected value.

---

[13] Johnston, L., Jr. The immediate and short-term effects of group discussion and individual study on risk-level. Unpublished doctoral dissertation, Florida State University, 1968.

However, these results may still be consistent with the rationality hypothesis since it is possible that subjects were maximizing subjective expected utility rather than expected value. Close examination of the four studies that produced evidence against the expected value hypothesis reveals that in each of these studies either the utilities could well have been different from the explicit values (Bem et al., 1965; Pruitt & Teger, 1969) or the subjective probabilities could have been different from the objective ones (Clausen, see Footnote 14; Wallach et al. 1964). The latter experiment was reviewed by Clark (1971). Its results are also considered to be the strongest evidence in support of the diffusion-of-responsibility hypothesis, yet they lend themselves to different interpretations. The risk-taking instrument used in the study consisted of items from the College Board Entrance Examination. Risk level was measured by the difficulty of the item, determined by the percentage of students in the population who answered the item correctly. Subjects' subjective perception of the difficulty of the item was not measured. It can be argued that group discussion increases subjects' confidence about their ability to answer the item correctly, and influences them to take what objectively is a more difficult item. Alternatively, since the group can select its ablest member to tackle a problem, it could also choose a more difficult item (i.e., one with lower odds of success). Clausen (see Footnote 14) utilized the same instrument in her experiment and indeed found a positive correlation between a subject's riskiness, that is, choice of a difficult item, and his confidence. It seems likely, therefore, that what was manipulated in Wallach et al.'s (1964) experiment was not shared responsibility but confidence and subjective perception of the difficulty of the items. In conclusion, a rationality hypothesis that adopts maximization of subjective expected utility as a criterion of rational choice is still a viable explanation for risky as well as for conservative shifts and should be subjected to direct test.

A study by Stoner (1968), though not a direct test of a subjective expected utility hypothesis, suggests a way of doing it and provides results that certainly justify the effort. In his study, subjects were presented with 18 phrases, selected by the experimenter, describing possible outcomes involved in 12 choice-dilemma items. Each phrase described one possible outcome. Subjects were asked to rank the 18 phrases "in the order in which they are important to you." (Although in the study this instrument is called a "value-ranking instrument," it is clearly a *utility*-ranking instrument.) In addition, subjects were asked to make the usual decisions on each item individually, and were subjected to the usual procedure of group discussion to consensus. The results show that for items producing a risky shift, the successful outcome of the risky alternative was ranked higher in importance than the outcome of the cautious, more certain alternative. The converse held for the items that produced cautious shifts. The results of this experiment may be misleading because the value-ranking instrument was *not* constructed by a valid objective procedure and (perhaps as a result) did not include phrases for all possible outcomes. Consequently, the entire experiment is open to serious criticism even concerning its immediate purpose, which was to test the value hypothesis. For a direct test of the subjective expected utility hypothesis, the utilities of all outcomes should be adequately represented, and properly dealt with as prescribed by a prediction model. It is argued later that a great deal of important information concerning the nature of the group process could be obtained by testing the subjective expected utility hypothesis.

---

[14] Clausen, G. S. Risk taking in small groups. Unpublished doctoral dissertation, University of Michigan, 1965.

INTERACTIVE HYPOTHESES

The affective and cognitive hypotheses do not consider the interaction among group members to be necessary or sufficient for shift; it is only a condition under which diffusion of responsibility or information exchange occurs. Other appropriate conditions may substitute for it. In contrast, interactive hypotheses conceive of the interaction as a necessary condition for shifts to occur. They are concerned more with identifying the social processes that produce the effect (e.g., disproportionate exercise of influence by some group members) than with the psychological mechanisms involved (e.g., affective or cognitive), and as such they constitute a class in themselves. At the same time, the interactive hypotheses differ from each other. Each assumes that shifts occur because some members exert more influence in the group decision making than others. But they disagree as to the reasons for this differential influence. In other words, their differences lie in what is assumed to be the basis of the influence, for instance, some personality characteristics, majority influence, or greater confidence in certain positions. Of course, since they make no explicit assumptions about psychological mechanisms, these theories are not necessarily incompatible with the affective or cognitive hypotheses.

Risk Takers as Leaders

The empirical evidence for and against this hypothesis was fully reviewed by Clark (1971). On the theoretical level the leadership hypothesis assumed the existence of stable personality characteristics of leadership and risk-taking tendency. Most of the evidence indicates that this assumption is invalid. The failure to find common and stable leadership characteristics (Gibb, 1969) is as glaring as the repeated failure to find convergence among risk-taking measures (regarding the lack of stable generalized risk-taking tendency see Goodman[15]; Kogan & Wallach, 1967a; Slovic, 1962, 1964; Weinstein, 1969). Adding to this the negative results of studies that tested the hypothesis itself, it appears that the leadership hypothesis may safely be discarded.

Extremity and Influence

This hypothesis views the greater influence exerted by some group members, not as a function of some stable personality characteristic (such as assumed by the leadership hypothesis), but as a function of the extreme position they take with regard to an issue. The same person, according to this hypothesis, may take an extreme position on one issue and influence the moderate group members, but then take a moderate position on another issue and himself be influenced by the more extreme members of the group. Extreme position is not limited to the choice of high risk levels (i.e., low probability of success) but includes low risk levels (i.e., high probability of success) as well. It is, therefore, a hypothesis that could account for group shifts toward greater risk taking as well as for shifts toward conservatism.

The extremity hypothesis was proposed by Burns (see Footnote 4) as a model to predict group shifts toward greater risk and conservatism. The amount and direction of shift is determined in the model by a regression equation. Burns argued that subjects respond on the probability scale as a Likert-type preference scale, so that choice of a very low probability of success indicates very intense and strong preference for the risky alternative; choice of a very high probability of success indicates very intense and strong preference for the more certain conservative alternative. He also argued that as "intensity of attitude varies with attitude extremity is well documented by past research [p. 38]," it is the group members who endorsed extreme positions (i.e., either very high or very low probabilities), that are more confident and committed to their position. Thus, the greater exercise of influence during the group interaction is attributed to the holder of these extreme positions. Two similar models were tested using choice-dilemma items on 180 group decisions and predicted the exact group decisions on 49% and 53% of the cases, respectively. This is far better than any other model that has been proposed, including a convergent averaging model that uses the mean of the initial risk-level choices as the predictor. The latter model could predict exactly only 36% of the group decisions. It was also shown that greater confidence was indeed associated with extreme positions.

---

[15] Goodman, B. Risky decisions by individuals and groups. Unpublished doctoral dissertation, University of Michigan, 1970.

Similar models focusing on extremity were also proposed by Myers, Murdock, and Smith.[16] The basis for their prediction stems from the congruity model of attitude change (Osgood, Suci, & Tannenbaum, 1957), which assumes that the resistance to change of an attitude object is proportional to its extremity or polarity. Again, these models were tested using choice dilemma items, and they predicted group decisions significantly better than the averaging model. The inadequacy of these hypotheses is discussed later, with a critical evaluation of the interactive hypotheses.

Group Problem-Solving Analogy: Commitment-to-Risk and Influence-through-Commitment Hypotheses

These hypotheses make the same predictions as the extremity hypothesis. The difference lies in the exact nature of the mechanism proposed to account for disproportionate influence. They all agree that confidence and commitment constitute this mechanism. Yet, the extremity hypothesis argues that confidence is the natural concomitant of taking an extreme position on the scale, since it reflects intensity, whereas the hypotheses reviewed in the present section suggest different causes for the greater confidence. They are based on an analogy from group problem-solving studies and several ad hoc assumptions.

Burnstein (1969) pointed out the relevance to group decision making involving risk of results coming from studies of individual versus group problem solving. These studies are concerned with whether group decisions are more *correct* than the average individual decision, and if so, what accounts for this superiority. Early studies by Gurnee (1937a, 1937b) and by Thorndike (1938a, 1938b) demonstrated the superiority of the group and pinpointed that "correct" members were more confident in their choices. The inference is made that by virtue of their greater confidence the "correct" members were also more committed to their choice and were more influential throughout the group discussion. Burnstein offered the following conjectural explanations:

> Perhaps increased commitment or confidence is an inherent consequence of choosing an alternative which is more likely to lead to failure. The conjecture would be, first, that the effort, in the form of extensive and anxious calculation, which precedes a decision is greater in the case of a risky choice than in the case of a conservative choice and, second, the unpleasant after-effect of choice, a post-decision dissonance, is greater following a risky choice than following a conservative one. If either or both of these suppositions are correct, the risky decision is likely to take on added attraction which could well correspond to a feeling of commitment to the choice and a sense of confidence in its wisdom. Let us call this the "commitment to risk" hypothesis. At the same time, a conservative shift should occur when members who initially prefer an alternative with a high probability of success are highly confident in the wisdom of their choice (a relatively infrequent conjunction of traits according to this analysis) and engage in decision making with others who initially prefer a risky alternative but are not confident of their choice. This we will label the "influence through commitment hypothesis [p. 384]."

These conjectures have been tested in research by Clausen (see Footnote 14) using College Board problems as a risk-taking instrument. In general, her research generated supportive evidence for the commitment-to-risk hypothesis. Greater confidence was found to be correlated with a choice of a more difficult (i.e., "riskier") problem, particularly for problems in the areas of mathematics, spatial relations, and sentence completion, where a risky shift in the Wallach et al. (1964) study was produced. In problems dealing with antonyms and analogies, confidence was not correlated with the difficulty of the problem, and the risky shift produced was relatively small. The influence-through-commitment hypothesis was not supported, because no conservative shifts were observed—even when greater commitment and confidence were displayed by the more conservative group members. It must be noted, however, that conclusions based on these results are limited to studies using College Board problems as a risk-taking instrument. Studies using choice dilemmas showed some differences: Greater confidence was correlated with extreme choices ($r = .29$, $p < .001$), both very risky and very conservative ones, and resulted in risky shifts as well as in conservative shifts (Burns, see Footnote 4).

---

[16] Myers, D. G., Murdoch, P., and Smith, G. F. Responsibility diffusion, drive enhancement, extremity influence, and information exchange effects on risky shift. Unpublished manuscript, Hope College, Holland, Michigan, 1969.

In general, the evidence that supports the extremity hypothesis also supports the commitment-to-risk and influence-through-commitment hypotheses. There is not much that could be said about evidence differentiating these two hypotheses from the extremity hypothesis. Testing the assumption underlying the correlation between extremity and confidence would provide this evidence. There are, however, important results that are not explained by any of the interactive hypotheses. These results indicate that a risky shift can be produced by individual subjects who merely listen to group discussion of the choice-dilemmas questionnaire but do not interact in any way (Kogan & Wallach, 1967b; Lamm, 1967). Though the shift is not of the usual magnitude, it is nevertheless substantial and significant. Also unaccounted for are the results showing that the group decisions are accepted by the subjects covertly as their private decisions (Lamm, 1967; Wallach & Kogan, 1965) and that this acceptance endures for a considerable time (Johnston, see Footnote 13; Wallach et al., 1962). Since all the interactive hypotheses assume that the exercise of disproportionate influence is critical, they cannot account for the risky shift produced without interaction, unless they suppose that the influence during interaction is informational rather than normative (see Deutsch & Gerard, 1955). The basis for the last statement is made clear in the discussion that follows.

STATISTICAL HYPOTHESES

According to the statistical models, a risky shift as well as a conservative shift is an artifactual phenomenon. It does not reflect a change of the individuals' risk-level preferences but appears as an inevitable outcome of an a priori rule of ordering and reconciling the individuals' preferences into one decision. Consequently, no assumptions concerning the exercise of influence or other affective and cognitive processes are made. The important variables that account for group shifts are only the group decision rule and the distribution of initial risk-level choices. The models differ in what they assume about these two variables.

Two decision-making models considered as alternative explanations of the risky shift have been reviewed by Burnstein (1969). One model is applicable when individuals' rank orders of preferences or judgments are secured and can be compared (correlated) with a "true" or criterion rank order. The other model requires that the rank-order data be processed according to a majority-rule decision scheme or, alternatively, the processing of the single most preferred judgments of the individuals according to the same scheme. In the latter case, the mode (or median) judgment is taken as the pooled group judgment. These models are based on a test theory model (Eysenck, 1939; Preston, 1938; Zajonc, 1966). In brief, the basic model suggests that if there is an initial tendency toward accuracy—better than 50% chance of being correct, or a positive correlation between judgment and criterion—then the correlation between pooled group judgment (average rank order across subjects) and the criterion will be greater than the average individual correlation. In other words, pooled group judgments will be better than the average individual judgments *only if* (*a*) the majority of the individual judgments tend to be correct (i.e., exhibit some positive correlation with the criterion); or (*b*) the majority of judgments are correct and the mode judgment is taken as the pooled group judgment. If riskiness (or conservatism) is substituted for correctness, then shift toward risk (and conservatism) follows directly from the model. When the average individual choice tends to be risky, a risky shift is predicted, and when it tends to be conservative, a conservative shift is predicted. Calculations were attempted on contrived data and yielded results in accordance with the model (see Burnstein, 1969; Zajonc, 1966). Luce and Raiffa (1957, pp. 358–359) have suggested another procedure for aggregating individual choices to predict group decisions according to the majority-rule model. This procedure was attempted on the contrived data and again yielded results in support of the model (Burnstein, 1969).

What could the results of studies producing shifts tell us about the majority-rule model as a valid explanation for group shifts? For one thing, to validate the model they would have to show that on items that produced a risky shift, the average individual tended to be risky (i.e., the majority of individuals made risky choices), and on items producing a conservative shift, the average individual tended to be conservative. In studies using College Board problems where a risky shift was found, the average of the initial probabilities chosen was quite low (.30), indicating strong initial tendency for risk (Clausen, see Footnote 14; Wallach et al, 1964). In two other studies, one with real physical consequences for the subjects (Bem et al., 1965) and the other with monetary payoffs (Pruitt & Teger, 1969), the average initial probabilities were .34 and .10, respectively. In both studies, a risky shift was produced. Thus, these results support the model as an explanation for the risky shift.

Results from studies that used the choice-dilemmas questionnaire are not quite consistent with the model. According to data of six studies reported by Pruitt and Teger (see Footnote 12), the average initial probability for the 10 items that produced a risky shift is .54; for the four most risky items, it is only .49, which is incongruent with the model; and for the two items that produced a conservative shift, the average is .75, congruent with the model. More congruent results for the model can be found in Stoner's (1968) study using a different set of choice-dilemma items. The average initial probability for six risky items is .42, and for four conservative items, the average is .76.

The results of the studies that used decision tasks involving real consequences are congruent with the majority-rule decision-making model. The same is true for the results of Stoner's (1968) study using a revised set of choice-dilemma items. Nevertheless, the results of the more commonly used choice-dilemmas questionnaire are incongruent with the model. Thus, the evidence with regard to the adequacy of the model as an explanation for the risky shift is inconclusive. Additional analysis is needed to determine whether the model can be considered a valid explanation for the risky shift.

If the majority-rule model has any validity in explaining shifts, positively skewed distributions of initial choices (i.e., with a risky majority) should yield a risky shift, and negatively skewed distributions (i.e., with a rather cautious majority) should yield a conservative shift, regardless of the means of these distributions. This could also be viewed as a simple conformity situation. Vinokur (1969) manipulated initial distributions using the six most risky items of the choice-dilemmas questionnaire. Though the positively skewed distributions yielded a significant risky shift (−.86; $p < .001$), the negatively skewed distributions did not yield a conservative shift, but instead yielded a small, nonsignificant risky shift (−.33), even though the mean of the negatively skewed distributions was .52 (i.e., on the conservative side of the scale). Furthermore, it was shown—using reanalyzed raw data from three other studies that had yielded risky shifts on the choice-dilemmas questionnaire (Wallach & Kogan, 1965; Wallach, Kogan, & Burt, 1965; Wallach et al., 1962)—that contrary to what would be expected, the percentage of positively skewed groups was virtually the same as the percentage of negatively skewed groups (42.2% and 43.6%, respectively; the rest were symmetrical). Additional analysis of the same raw data produced further evidence against the majority-rule model. Both the positively and negatively skewed distributions on the three most risky items of the choice-dilemmas questionnaire (4, 6, 7) with nearly the same means (.42, .45) yielded risky shifts with nearly the same magnitude (−1.58, −1.41). It was only on two nonrisky items (2, 5) that the positively and negatively skewed distributions yielded diametrically opposed shifts (−.50; $p < .05$; .53, $p < .01$) as predicted by the majority-rule model. Zajonc et al. (1968) examined data produced in their experiment comparing individual and group risk taking in a two-choice situation in light of various group decision models. None of these models, including the majority-rule model, could account for the group shifts toward conservatism produced in their experiment.

In view of all the evidence, the majority-rule model is untenable. Hence, group shifts cannot be considered as statistical artifacts of the way individual preferences are reconciled. Besides, as was already shown, there is a growing body of evidence indicating that the change in individuals' risk-level preferences is genuine. The individuals change their risk-level preferences following a discussion, without being asked to make any group decision (Wallach & Kogan, 1965) or even after listening to real group discussions of the choice-dilemma items (Kogan & Wallach, 1967b; Lamm, 1967). When they do reach group decisions, they accept the group decisions as their private risk-level preferences (Wallach & Kogan, 1965) and maintain their new risk-level preferences for at least 6 weeks (Johnston, see Footnote 13; Wallach et al., 1962). In general, the statistical models seem inadequate because they attempt to avoid assumptions about group processes that can be shown to be critical for a shift.

CONCLUSIONS AND DISCUSSION

This review suggests that of all the explanations for shifts toward greater risk or conservatism, only three are viable. These are the value (information relevant to the task) hypothesis, the extremity hypothesis, and the rationality (subjective expected utility) hypothesis. The first two hypotheses are supported by direct evidence and are not inconsistent with evidence from other studies. The third one, rationality (subjective expected utility), is consistent with the results of most studies reviewed here, but has not yet been subjected to direct test.

The two cognitive hypotheses, rationality (subjective expected utility) and value (information relevant to the task) could complement each other. Both argue that the crucial factor responsible for individual as well as group shifts toward greater risk (or caution) is *information relevant to the task* that is produced during the group discussion. While the rationality hypothesis explains how the same information is processed and acted upon, the value hypothesis focuses on how and why new information is produced.

The extremity hypothesis is clearly incompatible with the other two hypotheses. It differs in what it considers to constitute the crucial factor, the key mechanism, responsible for the shifts. Whereas the value and rationality hypotheses consider the information relevant to the issue to be the crucial factor, the extremity hypothesis considers the greater preference for a certain course of action, as reflected by the extreme position, to be the crucial one. It argues that the greater preference results in greater confidence and commitment to the extreme position and leads to greater influence. To put it differently, the two cognitive hypotheses can be contrasted with the extremity hypothesis in respect to the kind of influence that induces the shifts. The value and rationality hypotheses assume some kind of informational influence; the extremity hypothesis assumes normative influence (see Deutsch & Gerard, 1955). Most of the research evidence indicates it is informational influence that induces changes in risk-level preferences. We have already noted that shifts occur after discussion without group decision (Clark & Willems, 1969; Kogan & Wallach, 1967b; Lamm, 1967; Wallach & Kogan, 1965), as well as by mere exposure (listening) to group discussions (Kogan & Wallach, 1967b; Lamm, 1967), and that the acceptance of the new risk level endorsed by the group is covert as well as overt (Lamm, 1967; Wallach & Kogan, 1965) and relatively permanent (Johnston, see Footnote 13; Wallach et al., 1962). Shifts on nonrisky (neutral) choice dilemmas were shown to be the result of majority influence—a risky shift resulted from a risky majority, and a conservative shift resulted from a conservative majority. However, the risky shifts on the risky choice-dilemma items were shown to be unaffected by the existence of either a risky or a conservative majority; consequently, they can only be interpreted as the outcome of informational influence (Vinokur, 1969).

However, the extremity hypothesis—in fact, all the interactive hypotheses—demonstrated correlations between greater confidence and extreme position, and these results were marshaled as supportive evidence. It seems that these results can be viewed as equally compatible with the value and rationality hypotheses. We only have to assume that the more information and persuasive arguments favoring a certain alternative a subject may have, the more extreme his position (i.e., risk-level preference) will be. In other words, the extreme position taken by a subject already reflects the level and quality of the information he holds; the better that information, the more confident he will be in holding his position. It is analogous to the common finding in group problem-solving studies mentioned earlier, that the correct group member is also more confident. Thus, the greater confidence is not something inherent in holding an extreme position as the extremity hypothesis holds. Both confidence and extreme position are more often the result of the availability of better information and more persuasive arguments. The correlation between them may indeed drop to zero in ambiguous situations where good information is difficult to obtain, and persuasive arguments are harder to generate. Clausen's (see Footnote 14) study illustrates this point cogently. In her male groups, she found positive correlations between confidence and preferred risk level (i.e., item difficulty), for items dealing with mathematics, spatial relations, and sentence construction that produced large risky shifts in Wallach et al.'s (1964) study. She obtained zero-order correlations for items dealing with antonyms and analogies where virtually no shifts were found. Obviously, it is easier to demonstrate to others the correctness of one's solution to a mathematical problem than the appropriateness of one's choice of an antonym or certain analogy.

Because it is an interactive hypothesis, the extremity hypothesis cannot account for shifts that are produced by mere exposure to group discussion through listening. The cognitive value and rationality hypotheses, although explaining shifts in these circumstances, would predict a shift of the same magnitude when subjects participate in the group discussion themselves. However, the results show that the magnitude of these shifts is not nearly as great as that produced in interaction groups (Kogan & Wallach, 1967b; Lamm, 1967). It would seem that neither a cognitive nor an interactive process can, by itself, constitute a sufficient condition for the appearance of the phenomenon in its full strength. Therefore, it would be worthwhile to consider an explanation that includes both processes. The one we propose here is based on two related models: the partially shared information model and the group problem-solving model. The first model was originally proposed by Steiner (1966) to account for the relationship between group size and potential group productivity. The implications of the partially shared information model are consistent with results from studies concerned with group problem-solving models (Lorge & Solomon, 1955; Restle & Davis, 1962) and more general studies comparing individual and group performances (Barnlund, 1959; Hall, Mouton, & Blake, 1963; Thorndike, 1938a, 1938b, for comprehensive reviews see Collins & Guetzkow, 1964; Kelly & Thibaut, 1969; Lorge, Fox, Davitz, & Brenner 1958). These studies demonstrated the superiority of groups over the average individual on a variety of tasks, particularly when full interaction was allowed.

For any choice-dilemma problem there exists a pool of persuasive arguments for each alternative course of action. A persuasive argument can be equated with a correct solution in a group problem-solving situation. Once it has been brought up, it is likely to be accepted by the subjects as valid. The pool of persuasive arguments related to a given choice-dilemma item can reliably be obtained from a large population of subjects; thus, predictions concerning individual and group shifts could be made in advance. It is assumed that a subject reading a choice-dilemma item attends to the information contained in the item, makes some inferences, and organizes the information into arguments for and against the risky and the cautious alternative. Then, he reaches a decision concerning his preferred risk level on the basis of these arguments. If he comes up with many persuasive arguments favoring a particular alternative (i.e., the risky or the cautious one), he is likely to take an extreme position (i.e., either very high or very low risk level, depending on which alternative the arguments favor), and to be very confident and committed to his decision. His extreme position and greater confidence are *both* consequences of the quality and number of arguments he possesses. Furthermore, he may think his risk level is somewhat more extreme (i.e., either more risky or more conservative) than that taken by most others (evidence for this was presented above) because he feels that he possesses arguments that others may not be aware of.

If for a given choice dilemma there is a preponderance of persuasive arguments for one alternative, then individual and group shifts in risk level are predicted to occur in the direction of the alternative favored by the arguments. The strength of the shifts is first determined by the probability that an average subject will possess all or most of the persuasive arguments. If this probability is either very high or very low, *strong* shifts will not occur. If the probability is high, the individuals have already made their initial decision on the basis of the persuasive arguments. Thus, prior to discussion the group members are already extreme. If the probability is very low, it is unlikely that individuals will come up with those persuasive arguments, even after considerable thought. Therefore, it is also unlikely that a group will contain individuals who possess these persuasive arguments and use them in discussion. For these reasons, the strongest shifts are predicted on choice items with a preponderance of persuasive arguments for one alternative, *but* with a moderate probability that an average individual will possess them. If group shift toward greater risk (or conservatism) following persuasive arguments is substituted for group movement toward better problem solution, then the analogy between the risky shift and the group problem-solving model becomes apparent.

Given such an item with a preponderance of persuasive arguments to one direction, the strength of the shift is also determined by the conditions preceding the group decision. Very small shifts are predicted when the subject familiarizes himself with an item. He is likely to hit upon some but not all of the persuasive arguments during the process of familiarization. These shifts may be difficult to detect experimentally, which may explain the failures to replicate Bateson's (1966) and Flanders and Thistlethwaite's (1967) results. Moderate shifts are predicted for a situation where subjects merely listen to group discussion. Here the likelihood that persuasive arguments will be brought up during the group discussion is considerably increased. This will influence listeners who did not possess these arguments to change their decision. The strongest shifts are predicted for a situation where the subjects actually engage in group discussion, because the less extreme members can ask the more extreme members direct questions and clarify moot points.

If this explanation is valid, there should be no difference in the magnitude of the shift between groups forced to reach a unanimous group decision and groups in which individuals are asked to reach private decisions following discussion. Indeed, equally strong shifts were found in the two conditions by Wallach and Kogan (1965) and Lamm (1967). Furthermore, the strength of the shifts is predicted to increase with group size. The more members in the group, the more likely it is that the group will contain at least one person with the persuasive arguments; consequently, the more likely it is that the persuasive arguments will be brought up during discussion. A study by Teger and Pruitt (1967) demonstrated a considerable increase in the magnitude of the risky shift with increases in group size from 3 members (shift of $-.14$) to 4 ($-.59$) to 5 ($-.97$) for the 10 risky items of the choice-dilemmas questionnaire; and, likewise, an increase in conservative shift in groups of Size 3 ($-.05$), 4 (.12), 5 (.38) for the two conservative items of the choice-dilemmas questionnaire. Equivalent results are reported by Vidmar and Burdeny,[17] using a different set of choice items, for an increase of both the risky shift on risky items and conservative shift on conservative items with groups of Size 2, 3, 4, and 5 members. Similarly, the strength of the shift is predicted to increase with the heterogeneity (variance) of group members' initial choices. Again, a study by Vidmar (1970) showed significantly greater risky shifts in heterogenously composed groups than in homogenously composed groups.

So far, nothing has been assumed about the nature of the persuasive arguments. It remains to be seen whether the persuasive arguments are those appealing to cultural values of riskiness and caution in given situations, as Brown's value hypothesis implies, or whether they are merely informational arguments bearing more specifically upon the desirabilities (utilities) of the various possible outcomes in each situation.

Thus, the present review is in basic agreement with the conclusions reached by Clark (1971) concerning the necessary and sufficient conditions for the occurrence of the risky-shift phenomenon. The explanation based on partially shared information and group problem-solving models introduced above provides further clarification of those conditions, because it accounts for their structural relationships. According to this explanation, persuasive arguments raised in the discussion (Clark's Point *a*—"risk relevant content") are the primary cause. Clark's other conditions—(*b*) persons who perceive themselves to be relatively risky (or conservative for conservative items),

---

[17] Vidmar, N., and Burdeny, T. C. Interaction effects of group size and relative risk position with item type in the "group shift" effect. (Research Bulletin No. 128) London, Canada: University of Western Ontario, Department of Psychology, 1969.

(*c*) heterogeneity of initial opinions, and (*d*) severity of consequences of failure—are merely related to Point *a*, being the conditions under which Point *a* is operative. All these conditions appear jointly when the shift phenomenon occurs, and it is drastically reduced when even one is eliminated. The "risk-oriented instructions" specified by Clark (i.e., his Point *e*) are indeed necessary for the risky shift to occur, but the explanation of this in terms of "demand characteristics" is questionable, as is later discussed.

It has recently been argued that individual and group shifts toward greater risk or conservatism are not content-bound exceptions to the averaging tendency of the group, but represent a more general phenomenon: the polarization effect due to normative commitment (Moscovici & Zavalloni, 1969). It was demonstrated that following group discussion on opinion and judgmental items that were not risk related, individuals and groups shifted and endorsed more extreme positions. It is important to note that the explanation for risky and conservative shifts offered above was based on a general model that describes this effect of partially shared information on group productivity and does not view these shifts as content-bound exceptions. However, contrary to Moscovici and Zavalloni's view that these shifts are due mainly to normative or, in our terms, affective processes, the explanation presented here suggests that shifts are due to a cognitive or informational influence.

THEORETICAL EVALUATION AND IMPLICATIONS FOR RESEARCH

The conceptual bases for most of the hypotheses reviewed are drawn from the general area of group dynamics, with occasional reliance on behavioral decision theory. This reliance has never gone beyond the rejection of the rationality (expected value) hypothesis. It is argued here that both behavioral decision theory and theories of individual and group decision-making processes would benefit from the application of behavioral decision theory to complex situations (e.g., choice dilemmas).

Moreover, the conceptual difficulties concerning the definition of risk (Coombs & Huang, see Footnote 9), as well as the lack of convergent validity in the assessment of risk taking (Goodman, see Footnote 15; Slovic, 1962, 1964), raise serious doubts about whether studying group risk taking will increase our understanding of group processes. Instead of focusing on risk taking, it might be more fruitful if future studies concentrate on the decision process itself.

When seeking an understanding of how individual choices are processed and combined into some kind of group decision, knowledge of these individual choices is essential. Moreover, it is helpful to know how they are generated in the first place. Zajonc has suggested that this be a requirement for any standard group task (Zajonc, 1965). Obviously, it would be helpful to start with a good theory of decision making. Behavioral decision theory seems to be such a theory. It spells out relevant variables in any decision-making situation involving uncertainty, and suggests various relationships among them. The relevant variables in behavioral decision theory can be classified into the following three categories: (*a*) objective and subjective probabilities that certain events will occur and their outcomes follow; (*b*) values and utilities (i.e., subjective values) attached to those outcomes; (*c*) decision rules—ways of combining probabilities and values and utilities to arrive at a decision with respect to a choice of one from several alternative courses of action. The formulations of behavioral decision theory are rigorous, and this has made it appealing to experimentalists. Unfortunately, with very few exceptions (e.g., Brim, Glass, Lavin, & Goodman, 1962) experimental tests of the theory have been limited to a small range of experimental tasks and situations. It has usually consisted of outcomes with the same single value dimension—usually money—and subjects working for either real but small monetary payoffs or imaginary ones (cf. Becker & McClintock, 1967). More complex situations are those where the values of the various potential outcomes belong to different dimensions (e.g., money, prestige, power, health). According to behavioral decision theory, in order to make a decision in a complex situation, the subject must reconcile these different values and project their outcomes onto a single utility dimension. He must be able to compare the utilities of the potential outcomes.

The choice-dilemmas task seems to provide such a complex situation for testing behavioral decision theory. In turn, behavioral decision theory may be helpful in discovering the nature of the group decision-making process, which is the ultimate purpose of most of the studies reviewed here. How could this be successfully accomplished? First, it should be determined to what extent subjects attempt to maximize subjective expected utility. In terms of the risky-shift paradigm, it means that a subject's risk-level preference must reflect his estimation of the differences among the utilities associated with the potential outcomes in the choice-dilemma situation. A subjective expected utility model for predicting risk-level preferences for a typical choice-dilemma item can be constructed as follows:

Let U_c, U_s, U_f designate the utilities of the outcomes of the certain (cautious) alternative (U_c), of the risky alternative when successful (U_s), and when it fails (U_f); and p, q designate the *subjective* probabilities that the risky alternative will be successful (p) or unsuccessful (q). Assume, first, $U_s > U_c > U_f$; and, second, $p + q = 1$ (i.e., $q = 1 - p$) for every subject. Then:

$$\mathrm{SEU} = \sum_{i=1}^{n} p_i U_i, \qquad [1]$$

where SEU is the subjective expected utility. Subjective expected utility for the risky alternative:

$$\mathrm{SEU}_R = pU_s + (1 - p)U_f, \qquad [2]$$

and subjective expected utility of the certain alternative:

$$\mathrm{SEU}_C = 1 \cdot U_c. \qquad [3]$$

If maximization of subjective expected utility provides the decision rule, then the choice of the risky alternative should follow only when its subjective expected utility is equal to or greater than the subjective expected utility of the certain cautious alternative: that is, when

$$\mathrm{SEU}_R \geqq \mathrm{SEU}_C, \qquad [4]$$

or,

$$pU_s + (1 - p)U_f \geqslant 1 \cdot U_c. \qquad [5]$$

Consequently, the risky alternative should be chosen if and only if

$$p \geqq \frac{U_c - U_f}{U_s - U_f} \qquad [6]$$

In other words, the *lowest* probability of success that a subject is predicted to consider acceptable (i.e., his risk-level preference) in order to attempt the risky alternative if he uses maximization of subjective expected utility as a decision rule is:

$$\hat{p} = \frac{U_c - U_f}{U_s - U_f} \qquad [7]$$

($\hat{p}$: the ratio of the utilities of the outcomes, as judged by the subject, which also serves as a predictor of the subject's risk-level preference, p).

Several important consequences of the subjective expected utility model are indicated by Equation 6. First note that the p that maximizes subjective expected utility is not determined by the absolute utility of any single outcome alone, but by the relative differences between all the utilities. Therefore, Clark's (1971) conclusion that the risky shift depends on the utility of failure, U_f (i.e., "severity of the consequences of failure") is certainly compatible with the model. However, if the model proves to be valid, the conclusion will have to be extended to include the utilities of the other outcomes (i.e., U_c and U_s) as well.

Second, in order to reach a group decision concerning the lowest p, it is not necessary for the subjects to reconcile their absolute utilities directly, as is often assumed, but merely to agree on their relative magnitudes, that is, on the ratio among those utilities.

Third, for the occurrence of a shift toward risk, "risk-oriented instructions" (i.e., the standard instructions requesting subjects to choose the *lowest* probability that is acceptable) are necessary. Without the word "lowest" the meaning of the task becomes vague, since invariably any higher probability is preferred to a lower one—a point that may be raised in a group discussion inhibiting the shift toward risk. When "neutral" instructions that omit the word "lowest" are presented to subjects, the model would predict that the usual significant risky shift will not occur. Indeed this was found by Clark (1971). The model would also predict that "neutral instructions" will not reduce the conservative shift on conservative choice dilemmas and would possibly increase it—a matter for future research. Thus the necessity of risk-oriented instructions to produce the risky shift is shown by the subjective expected utility model (see Equation 6) and confirmed by Clark and Willems' (1969) data. However, his interpretation that these instructions impose "demand characteristics" and that the risky shift is the outcome of "demand characteristics" is questionable. From an empirical point of view, this interpretation does not account for conservative shifts that are produced with the same risk-oriented instructions. From a theoretical point of view, these instructions are necessary to produce a risky shift (but not a conservative shift) as is shown by Equation 6 of the subjective expected utility model.

If the subjective expected utility model proves to be valid for individual choice behavior—in other words, if subject's initial choices do reflect the relationships among utilities required by subjective expected utility theory—then it can be used to clarify the exact nature of the group process responsible for risky and conservative shifts. (That is, by observing how the group process affects the relationship among utilities the nature of the group process could be determined.) In order to validate the subjective expected utility model, positive correlations between p and $\hat{p}$ (i.e., $\hat{p}$ as in Equation 7) should be obtained for each choice dilemma across subjects, as well as within each subject across several choice dilemmas.

Before proceeding with the analysis of the group process, let us define some new terms. Let p_1, p_2, $\hat{p}_1$, and $\hat{p}_2$ designate the risk-level choice before (p_1) and after (p_2) group discussion, and the predicted probabilities based on the utilities' ratio also before ($\hat{p}_1$) and after ($\hat{p}_2$) group discussion.

Given data regarding p_1, p_2, $\hat{p}_1$, and $\hat{p}_2$, it is possible to test whether a cognitive or an affective process is responsible for the shifts. Furthermore, if a cognitive process underlies the shifts, its nature could easily be determined by a simple analysis. More specifically, it will be shown how the effects of two viable explanations for risky and conservative shifts could be differentiated. One explanation based on partially shared information and group problem-solving models contends that the shifts are produced by the introduction of new information (i.e., persuasive arguments) during the group discussion. This explanation is referred to as the informational influence hypothesis. In contrast, the other explanation suggests that the shifts toward risk and conservatism are not due to the influence of *new* information but to a process of rationalization in which better use is made of the same information according to the subjective expected utility model (i.e., better maximization of subjective expected utility). This explanation is referred to as the rationality subjective expected utility hypothesis.

According to the informational influence hypothesis, there has to be a change in $\hat{p}$ in the direction of the change in p as a result of the additional information relevant to the utilities on which $\hat{p}$ is based. However, there need not be any increase in the goodness of fit between p and $\hat{p}$ subsequent to the group discussion (i.e., the absolute difference between p and $\hat{p}$ need not decrease), as increased rationality during group discussion is not assumed by this hypothesis.

According to the rationality hypothesis, there should *not* be a change in the $\hat{p}$ subsequent to the group discussion, since no *new* information on the utilities has been added. The shifts in p are in the direction of the initial $\hat{p}$ because of subjects changing their risk-level choices to conform to their utility ratio in order to maximize subjective expected utility. Consequently, the goodness of fit between p and $\hat{p}$ should increase considerably after group discussion (i.e., the absolute difference between p and $\hat{p}$ should decrease).

If both informational influence and rationalization processes are operative, there will be both a change in $\hat{p}$ in the direction of the change in p and an increase in the goodness of fit between the two.

There remains, of course, the possibility that neither of these two cognitive hypotheses is responsible for the risky and conservative shifts. In this case there should not be a change in $\hat{p}$ after group discussion; instead, the goodness of fit between p and $\hat{p}$ should decrease. Evidence of this kind would suggest that an affective process is responsible for the shifts—that new utilities [e.g., utility for risk, or for appearing risky, as the risk-as-a-value (information about others' choices) hypothesis suggests] not represented in the outcomes of the choice dilemma (i.e., Equation 4) are influencing the risk-level choices. (Or, although unlikely, it may be that group members are moving away from subjective expected utility and applying a different decision rule.) A summary of this analysis is represented in Table 1.

Preliminary results of studies by the author and other researchers at the Research Center for Group Dynamics in the Institute for Social

TABLE 1

SUBJECTIVE EXPECTED UTILITY ANALYSIS OF THE EFFECTS OF GROUP PROCESSES ON INDIVIDUAL AND GROUP DECISIONS INVOLVING RISK

Choice-dilemma type	Change in p, risk preference	Analysis and results		Evidence in support of:
		Change in $\hat{p}$, utilities' ratio	Fit between p and $\hat{p}$	
Risky Conservative	$p_2 - p_1 < 0$ $p_2 - p_1 > 0$	$\hat{p}_2 - \hat{p}_1 < 0$ $\hat{p}_2 - \hat{p}_1 > 0$	$\lvert p_1 - \hat{p}_1 \rvert = \lvert p_2 - \hat{p}_2 \rvert$ $\lvert p_1 - p_1 \rvert = \lvert p_2 - \hat{p}_2 \rvert$	Informational influence hypothesis
Risky Conservative	$p_2 - p_1 < 0$ $p_2 - p_1 > 0$	$\hat{p}_2 = \hat{p}_1$ $\hat{p}_2 = \hat{p}_1$	$\lvert p_1 - \hat{p}_1 \rvert > \lvert p_2 - \hat{p}_2 \rvert$ $\lvert p_1 - p_1 \rvert > \lvert p_2 - p_2 \rvert$	Rationality (subjective expected utility) hypothesis
Risky Conservative	$p_2 - p_1 < 0$ $p_2 - p_1 > 0$	$\hat{p}_2 - \hat{p}_1 < 0$ $\hat{p}_2 - \hat{p}_1 > 0$	$\lvert p_1 - \hat{p}_1 \rvert > \lvert p_2 - \hat{p}_2 \rvert$[a] $\lvert p_1 - \hat{p}_1 \rvert > \lvert p_2 - p_2 \rvert$[a]	Informational influence and rationality (subjective expected utility) hypotheses
Risky Conservative	$p_2 - p_1 < 0$ $p_2 - p_1 > 0$	$\hat{p}_2 = \hat{p}_1$ $\hat{p}_2 = \hat{p}_1$	$\lvert p_1 - \hat{p}_1 \rvert < \lvert p_2 - \hat{p}_2 \rvert$ $\lvert p_1 - \hat{p}_1 \rvert < \lvert p_2 - p_2 \rvert$	Affective hypothesis (or change in decision rule)

[a] Note that the increase in goodness of fit can result from either changes in p or $\hat{p}$, or both. According to the rationality hypothesis, which purports to account for the shifts in p, the goodness of fit should increase as a result of the changes in p rather than in $\hat{p}$. Consequently, the results that satisfy the above mentioned equations constitute evidence for the rationality hypothesis *only* if the mean change in p is greater than the mean change in $\hat{p}$ (i.e., $\lvert p_2 - p_1 \rvert > \lvert \hat{p}_2 - \hat{p}_1 \rvert$).

Research offer strong support for the subjective expected utility model of individual choice behavior for the choice-dilemma problems. Furthermore, these results indicate that informational influence rather than rationalization is the process responsible for group shifts. However, more research, employing a greater variety of tasks, is needed before the effects of group processes on individual and group decisions involving risk will be fully understood.

REFERENCES

BARNLUND, D. C. A comparative study of individual, majority, and group judgment. *Journal of Abnormal and Social Psychology*, 1959, **58**, 55–60.

BATESON, N. Familiarization, group discussion and risk taking. *Journal of Experimental Social Psychology*, 1966, **2**, 119–129.

BECKER, G. M., & MCCLINTOCK, C. G. Value: Behavioral decision theory. *Annual Review of Psychology*, 1967, **18**, 239–286.

BEM, D. J., WALLACH, M. A., & KOGAN, N. Group decision making under risk and aversive consequences. *Journal of Personality and Social Psychology*, 1965, **1**, 453–460.

BLANK, A. D. Effects of group and individual conditions on choice behavior. *Journal of Personality and Social Psychology*, 1968, **8**, 294–298.

BRIM, O. G., GLASS, D. C., LAVIN, D. E., & GOODMAN, N. *Personality and decision processes.* Stanford: Stanford University Press, 1962.

BROWN, R. *Social psychology.* New York: Free Press, 1965.

BURNSTEIN, E. An analysis of group decisions involving risk. *Human Relations*, 1969, **22**, 381–395.

CLARK, R. D. Group-induced shift toward risk: A critical appraisal. *Psychological Bulletin*, 1971, **76**, 251–270.

CLARK, R. D., & WILLEMS, E. P. Where is the risky shift. *Journal of Personality and Social Psychology*, 1969, **13**, 215–221.

COLLINS, B. E., & GUETZKOW, H. A. *Social psychology of group processes for decision making.* New York: Wiley, 1964.

COOMBS, C. H., & PRUITT, D. C. Components of risk in decision making: Probability and variance preferences. *Journal of Experimental Psychology*, 1960, **60**, 265–277.

COOMBS, C. H., & PRUITT, D. C. Some characteristics of choice behavior in risky situations. *Annals of the New York Academy of Sciences*, 1961, **89**, 784–794.

DEUTSCH, M., & GERARD, H. G. A study of informational social influences upon individual judgment. *Journal of Abnormal and Social Psychology*, 1955, **51**, 629–636.

EDWARDS, W. Behavioral decision theory. *Annual Review of Psychology*, 1961, **12**, 473–498.

EDWARDS, W., LINDMAN, H., & PHILLIPS, L. D. Emerging technologies for making decisions. In, *New directions in psychology*. Vol. II. New York: Holt, Rinehart & Winston, 1965.

EYSENCK, H. J. The validity of judgments as a function of the number of judges. *Journal of Experimental Psychology*, 1939, **25**, 650–654.

FESTINGER, L. A theory of social comparison processes. *Human Relations*, 1954, **7**, 117–140.

FLANDERS, J. P., & THISTLETHWAITE, D. L. Effects of familiarization and group discussion upon risk taking. *Journal of Personality and Psychology*, 1967, **5**, 91–97.

GIBB, C. A. Leadership. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology.* (2nd ed.) Vol. 4. Cambridge, Mass.: Addison-Wesley, 1969.

GURNEE, H. Maze learning in the collective situation. *Journal of Psychology*, 1937, **3**, 437–444. (a)

GURNEE, H. A comparison of collective and individual judgments of fact. *Journal of Experimental Psychology*, 1937, **21**, 106–112. (b)

Hall, E., Mouton, J., & Blake, R. Group problem solving effectiveness under conditions of pooling vs. interaction. *Journal of Social Psychology*, 1963, **59**, 147–157.

Kelley, H. H., & Thibaut, J. W. Group problem solving. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology*. (2nd ed.) Vol. 4. Cambridge, Mass.: Addison-Wesely, 1969.

Kogan, N., & Wallach, M. A. Risk taking as a function of the situation, the person, and the group. In, *New directions in psychology*. Vol. 3. New York: Holt, Rinehart & Winston, 1967. (a)

Kogan, N., & Wallach, M. A. The risky-shift phenomenon in small decision-making groups: A test of the information-exchange hypothesis. *Journal of Experimental Social Psychology*, 1967, **3**, 75–85. (b)

Lamm, H. Will an observer advise higher risk-taking after hearing a discussion of the decision problem? *Journal of Personality and Social Psychology*, 1967, **6**, 467–471.

Levinger, G., & Schneider, D. J. A test of the risk is a value hypothesis. *Journal of Personality and Social Psychology*, 1969, **11**, 165–169.

Lonergan, B. G., & McClintock, C. G. Effects of group membership on risk-taking behavior. *Psychological Reports*, 1961, **8**, 447–455.

Lorge, I., & Solomon, H. Two models of group behavior in the solution of Eureka-type problems. *Psychometrika*, 1955, **20**, 139–148.

Lorge, I., Fox, D., Davitz, J., & Brenner, M. A survey of studies contrasting the quality of group performance and individual performance, *Psychological Bulletin*, 1958, **55**, 337–372.

Luce, R. D., & Raiffa, H. *Games and decisions: Introduction and critical survey*. New York: Wiley, 1957.

Madaras, G. R., & Bem, D. J. Risk and conservatism in group decision making. *Journal of Experimental Social Psychology*, 1968, **4**, 350–365.

Marquis, D. G., & Reitz, H. J. Uncertainty and risk taking in individual and group decisions. *Behavioral Science*, 1969, **14**, 281–288.

Moscovici, S., & Zavalloni, M. The group as a polarizer of attitudes. *Journal of Personality and Social Psychology*, 1969, **12**, 125–135.

Preston, M. G. A note on the reliability and validity of the group judgment. *Journal of Experimental Psychology*, 1938, **22**, 462–471.

Pruitt, D. G., & Teger, A. I. The risky shift in group betting. *Journal of Experimental Social Psychology*, 1969, **5**, 115–126.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. *The measurement of meaning*. Urbana: University of Illinois Press, 1957.

Rabow, J., Fowler, F. J., Bradford, D. L., Hofeller, M. A., & Shibuya, Y. The role of social norms and leadership in risk-taking. *Sociometry*, 1966, **29**, 16–27.

Restle, R., & Davis, J. H. Success and speed of problem solving by individuals and groups. *Psychological Review*, 1962, **69**, 520–536.

Slovic, P. Convergent validation of risk taking measures. *Journal of Abnormal and Social Psychology*, 1962, **65**, 68–71.

Slovic, P. Assessment of risk taking behavior. *Psychological Bulletin*, 1964, **51**, 220–233.

Steiner, I. Models for inferring relationships between group size and potential group productivity. *Behavioral Science*, 1966, **11**, 273–283.

Stoner, J. A. F. Risky and cautious shifts in group decisions. The influence of widely held values. *Journal of Experimental Social Psychology*, 1968, **4**, 442–459.

Teger, A. I., & Pruitt, D. G. Components of group risk taking. *Journal of Experimental Social Psychology*, 1967, **3**, 189–205.

Thorndike, R L. The effect of discussion upon the correctness of group decisions, when the factor of majority is allowed for. *Journal of Social Psychology*, 1938, **9**, 343–362. (a)

Thorndike, R. L. On what type of task will a group do well? *Journal of Abnormal and Social Psychology*, 1938, **33**, 409–413. (b)

Vidmar, N. Group composition and the risky shift. *Journal of Experimental Social Psychology*, 1970, **6**, 153–166.

Vinokur, A. Distribution of initial risk levels and group decisions involving risk. *Journal of Personality and Social Psychology*, 1969, **13**, 207–214.

Wallach, M. A., & Kogan, N. The roles of information, discussion and consensus in group risk taking. *Journal of Experimental Social Psychology*, 1965, **1**, 1–19.

Wallach, M. A., Kogan, N., & Bem, D. J. Group influence on individual risk taking. *Journal of Abnormal and Social Psychology*, 1962, **65**, 76–86.

Wallach, M. A., Kogan, N., & Bem, D. J. Diffusion of responsibility and level of risk taking in groups. *Journal of Abnormal and Social Psychology*, 1964, **68**, 263–274.

Wallach, M. A., Kogan, N., & Burt, R. B. Can group members recognize the effects of group discussion upon risk taking? *Journal of Experimental Social Psychology*, 1965, **1**, 379–395.

Wallach, M. A., & Wing, C. W., Jr. Is risk a value? *Journal of Personality and Social Psychology*, 1968, **9**, 101–106.

Weinstein, M. S. Achievement motivation and risk preference. *Journal of Personality and Social Psychology*, 1969, **13**, 153–172.

Willems, E. P. Risk is a value. *Psychological Reports*, 1969, **24**, 81–82.

Zajonc, R. B. The requirements and design of a standard group task. *Journal of Experimental Social Psychology*, 1965, **1**, 71–88.

Zajonc, R. B. *Social psychology: An experimental approach*. Belmont, Calif.: Wadsworth, 1966.

Zajonc, R. B., Wolosin, R. J., Wolosin, M. A., & Sherman, S. J. Individual and group risk taking in a two-choice situation. *Journal of Experimental Social Psychology*, 1968, **4**, 89–106.

Zajonc, R. B., Wolosin, R. J., Wolosin, M. A., & Sherman, S. J. Group risk-taking in a two-choice situation: Replication, extension, and a model. *Journal of Experimental Social Psychology*, 1969, **5**, 127–140.

(Received March 31, 1970)

Psychological Bulletin
1971, Vol. 76, No. 4, 251–270

GROUP-INDUCED SHIFT TOWARD RISK:

A CRITICAL APPRAISAL[1]

RUSSELL D. CLARK, III[2]
Florida State University

The present study critically examines the risky-shift phenomenon by reviewing the prominent theories and the empirical evidence confirming and disconfirming each. These theories are the familiarization hypothesis, the leadership hypothesis, the diffusion-of-responsibility hypothesis, and the risk-as-value hypothesis. The present evidence suggests that the occurrence of the phenomenon is dependent upon (*a*) risk-relevant content, (*b*) persons who perceive themselves to be relatively risky, (*c*) heterogeneity of initial opinions, (*d*) situations for which the consequences of failure are not severe, and (*e*) risk-oriented instructions. The risk-as-value hypothesis seems to account for most of these necessary conditions. However, just how the cultural value of risk is mediated and the relationship between risk and instructions is less clear. In addition, future research will have to determine whether modification of these five necessary conditions or the formulation of additional conditions is needed to account for the apparent finding that individuals, who initially deviate from others in a culturally valued direction, subsequently conform to manipulated norms that conflict with the initially elicited value. The very recent research raises troublesome and complicated questions in this regard.

It is often assumed that group arrangements for solving problems and making decisions suppress or dampen individual tendencies toward boldness, innovation, and daring. One set of circumstances for which this assumption no longer appears to hold is in decisions or choices that embody a clear dimension of risk. By now, a host of studies, initiated by Stoner[3] in 1961 and given great impetus by Wallach and Kogan, has demonstrated that, on the average, individuals advocate greater degrees of risk taking following participation in a group than they had previously advocated in private (Bem, Wallach, & Kogan, 1965; Blank, 1968; Kogan & Wallach, 1967a; Pruitt & Teger, 1969; Stoner, 1968; see Footnote 3; Teger & Pruitt, 1967; Wallach & Kogan, 1965; Wallach, Kogan, & Bem, 1962, 1964). The phenomenon occurs when subjects in experiments are asked to serve as advisers to hypothetical persons who are described on paper as being confronted by a choice dilemma (Kogan & Wallach, 1964; Stoner, see Footnote 3); the hypothetical persons must choose between an alternative that is less attractive but safe and a situation that is more attractive but less likely to succeed. An example is an electrical engineer who may remain at his current job at a modest but adequate salary or take a new job that offers a much higher potential income but no assurance of long-range security. Subjects in the typical experiment make their recommendations regarding the engineer's accepting the new job, on a scale of odds varying from 1 in 10 through 10 in 10 that the new job will succeed. They choose their recommendation in terms of the minimum odds of success they would find acceptable before recommending that the engineer take the new job. After making their recommendations for a whole set of such items, subjects meet in groups and discuss each item in open-ended discussion. After the discussion they make new individual recommendations for each item. The typical finding is that on the average, subjects shift toward greater risk; that is, they recommend accepting alternatives at lower odds of success after the group discussion than they did before. This is the prototype of the group-induced shift toward risk or, simply, the risky shift.

---

[1] This review was supported in part by Training Grant BTTBT 217 from the National Institute of Mental Health to the Graduate Training Program in Social Psychology, University of Kansas; and in part by Grant MH17605-01 from the National Institute of Mental Health to Edwin P. Willems. The author is indebted to Walter H. Crockett, Anthony J. Smith, and Edwin P. Willems for making numerous suggestions on previous drafts of this manuscript.

[2] Requests for reprints should be sent to Russell D. Clark, III, Department of Psychology, Florida State University, Tallahassee, Florida 32306.

[3] Stoner, J. A. F. A comparison of individual and group decisions involving risk. Unpublished master's thesis, School of Industrial Management, Massachusetts Institute of Technology, 1961.

This group-induced shift toward risk has been obtained with a wide range of decision content, including hypothetical decisions (Teger & Pruitt, 1967; Wallach & Kogan, 1965; Wallach et al., 1962) and situations where positive and negative consequences to subjects may actually result from their decisions (Bem et al., 1965; Blank, 1968; Pruitt & Teger, 1969; Wallach et al., 1964). The phenomenon has been found with graduate business students (Stoner, see Footnote 3), English students (Bateson, 1966), Israeli students (Rim, 1964a), and American students (Kogan & Wallach, 1967a), senior executives (Marquis, 1962), and males and females (Wallach et al., 1962). It occurs in groups who engage in direct, face-to-face interactions (Kogan & Wallach, 1967a), groups who simply exchange information (Clark & Willems, 1969; Teger & Pruitt, 1967), groups who communicate over an intercom system (Kogan & Wallach, 1967b), and groups who merely observe other groups interacting (Lamm, 1967).

Two important things stand out regarding the risky-shift phenomenon: the recency of its discovery and the fact that none of the most common social psychological theories explains it adequately. Social psychologists have long been interested in the differences between decisions arrived at through some group process and those that are the product of individuals' efforts (Jones & Gerard, 1967; Kelley & Thibaut, 1969). However, the prevailing views are either that group decisions will be more conservative than individual decisions or that group decisions will represent an averaging, or simple convergence, of individual decisions. Many experimental results can be cited to support each view.

An experiment by Schachter (1951), as well as others cited in Cartwright and Zander (1960, pp. 165–341), found that group members tend to exert the most influence attempts toward those group members who deviate at the extremes; as the opinion of a deviant member is perceived to move toward the group's central tendency, the amount of communication directed toward him decreases.

In support of the view that groups exert conservative influences on the individual, Barnlund (1959), comparing groups with individuals on drawing logical conclusions from prepared arguments, found that group members were more cautious and deliberate. He concluded: "the necessity of explaining a conclusion forced many students to be more self-critical [p. 58]"; and, "group discussion was found to stimulate more careful thinking, to lead to a consideration of a wider range of ideas, and to provoke more objective and critical testing of conclusions [pp. 59–60]." Schein (1965), in reviewing studies on group discussion, made a similar point. The group is more likely than the individual to identify errors of judgment before action is taken. Greater care and deliberation as a function of group influence is also reflected in the finding by Zander and Medow (1963) that subjects working in groups more often lowered their aspirations following an unexpectedly poor performance than did those working alone. Evidence from Allport (1924) and Farnsworth and Behner (1931) suggests that subjects tend to give more moderate judgments of weights and odors when in a group situation than when alone.

While a variety of studies indicate that groups make more conservative judgments than individuals, it should be emphasized that none of the above studies offers a systematic basis for distinguishing between group and individual risk taking. Few, if any, of the studies have been directly concerned with the dimension of risk.

Finally, the findings from studies of conformity can account for the reduction in groups of the initial variation among the responses of group members (Asch, 1958; Festinger, 1954; Krech, Crutchfield, & Ballachry, 1962; Sherif, 1935, 1936). However, as Brown (1965) pointed out, conformity cannot account for the most important feature of the risky-shift phenomenon: the systematic shift toward risk. While variability in decisions is reduced after discussion, individuals are more willing to accept risky decisions after participating in interacting groups than they had been initially.

Perhaps in the recognition that other hypotheses than conformity would have to be formulated to account for the risky shift, a variety of explanations have been proposed in recent years to explain the phenomenon. The four that have generated the greatest amounts of empirical research may be called the familiarization hypothesis, the leadership hypothesis, the diffusion-of-responsibility hypothesis, and the risk-as-value hypothesis. The evidence for each of these hypotheses is discussed in turn.

THE FAMILIARIZATION HYPOTHESIS

It might be argued that the risky shift is not a true group effect, but a pseudogroup effect, as defined by Secord and Backman (1964). Pseudogroup effects occur in groups, but do not actually result from group processes. They can be reproduced in individuals operating independently under appropriate circumstances. Two studies, one by Bateson (1966) and the other by Flanders and Thistlethwaite (1967), indicate that the risky-shift phenomenon might be attributed to a pseudogroup phenomenon.

Bateson argued that group discussion allows persons to become more familiar with the situations being discussed, and contended that this increased familiarity with the situations is responsible for the observed shift toward risk. According to this hypothesis, any procedure that will increase familiarity with an issue involving risk will cause persons to become more risky on that issue. Informal observation and anecdotal evidence provide some support for this hypothesis. For many animals in unfamiliar situations or humans in strange circumstances (e.g., boys at a strange swimming hole) initial periods of caution, exploration, and "feeling out" are usually followed by behaviors that are much more bold and daring. To test this general hypothesis, Bateson compared the riskiness of groups who discussed five of the choice dilemmas with that of individuals who merely wrote briefs on these items indicating the points in favor of and against the various alternatives. Results were consistent with his hypothesis. Both procedures resulted in risky shift that differed from zero, but the two procedures did not differ significantly in the amount of shifts toward risk. Since the group discussion appeared to be irrelevant, Bateson concluded that the risky shift may be a simple matter of familiarization with the choice dilemmas. However, Pruitt and Teger[4] conducted two similar studies and were unable to replicate Bateson's results. They found no evidence that familiarity with the items, by itself, produced a risky shift.

On the other hand, Flanders and Thistlethwaite (1967) did obtain results similar to Bateson's. After preliminary measurement on the hypothetical dilemmas, each subject was assigned at random to one of four conditions. Subjects in the first condition worked individually, familiarizing themselves with the items on the assumption that they were going to discuss them later with other subjects; they were asked to concentrate on the important issues, making lists of "pros" and "cons." In the second condition, groups of three subjects discussed the items, according to the typical group discussion procedure. To determine whether discussion would add effects not already induced by familiarization, subjects in a third condition both familiarized themselves with the items and participated in three-man discussions of them. Finally, in a control condition, subjects merely filled out the questionnaire twice. There was a significant shift toward risk in all conditions except the control condition, and the familiarization procedure alone produced as much risky shift as the discussion condition or the discussion-plus-familiarization condition. The authors concluded that group discussion adds nothing to the effects of familiarization with the items; that is, discussion is the occasion for familiarization, but it is not a necessary condition of the risky shift. In other words, by this account, the risky-shift phenomenon is a pseudogroup effect.

Kogan and Wallach (1967c), however, have questioned the results of the familiarization studies. They compared groups who discussed the items with groups who merely listened to tape recordings of the group discussions, a condition that may be regarded as familiarization procedure. Risky shifts were found in both conditions, but the shifts under the discussion condition were significantly larger than under the listening condition. They argued that similar information was available to subjects in both conditions, so that if familiarization was sufficient to account for the risky shift, there should be no difference between the conditions. Since a difference was found, the authors felt that an explanation involving the vicarious experiencing of the group interaction is needed to account for the risky-shift phenomenon.

---

[4] Pruitt, D. G., and Teger, A. I. Is there a shift toward risk in group discussion? If so, is it a group phenomenon? If so, what causes it? Paper presented at the annual meeting of the American Psychological Association, Washington, D. C., September 1967.

The difference in results between these two experiments may be due to subtle differences in procedure. In their familiarization instructions, Flanders and Thistlethwaite created the expectation in their subjects that they were to engage in a group discussion later on. Part of their instructions read, "We want each of you to be prepared to discuss the choices to be made so that you will not have to spend group discussion time restudying the problem [p. 93]." In contrast, Kogan and Wallach did not lead their subjects to expect that a group discussion would follow the familiarization procedure. To test the effects of differences in instructions, Teger, Pruitt, St. Jean, and Haaland (1970) compared two familiarization conditions. In one condition, subjects wrote out the pros and cons of the various dilemmas items under the impression that they were subsequently to make new choices that would be revealed in a group discussion. In the other condition, subjects also wrote out the pros and cons of the various items, but they were under the impression that their subsequent choices would remain private. Neither familiarization condition produced significant shifts toward risk, but the traditional group discussion conditions did. Teger et al. have conducted additional tests of the familiarization hypothesis with nonsupporting results.

In the light of this evidence, it seems fair to conclude that the familiarization hypothesis cannot account for the risky-shift phenomenon. Blank's findings (1968) add further weight to this conclusion. In this experiment, subjects placed bets for upcoming throws of dice over many trials. Bets were selected from a table of odds on which outcomes with high probabilities had low winnings attached, and outcomes with low probabilities had high winnings attached. Subjects either played the betting game alone or in the simple, noninteractive presence of others. Since all subjects were exposed to the same materials and similar choices, there should have been no differences in degree of familiarization. And yet, over trials, subjects who played in the presence of others came to place more low-probability bets than subjects who played alone.

The Leadership Hypothesis

According to the second hypothesis, risk takers are perceived as group leaders.[5] Those who are initially more inclined to take risks also tend to be more dominant and influential in the group discussion; hence, the risky shift can be accounted for in terms of the influence of risky leaders. Marquis (1962) and Collins and Guetzkow (1964) have argued that this is the most likely explanation of the phenomenon.

In one test of this hypothesis, Wallach et al. (1962) asked group members to rank each other for degree of influence exerted in the group discussion. There was general agreement as to who had been most influential. In support of the hypothesis, group members with high initial levels of risk taking were perceived as more influential in the discussion. However, this study is not an adequate test of the hypothesis, since the outcome may simply reflect the fact that the group ends up at a riskier position, prompting group members to attribute greater influence to those who initially favored higher risk taking, regardless of whether they were actually more influential.

It is not necessary that the group reach a consensus in order for the relationship between initial level of risk and persuasiveness to hold. Wallach, Kogan, and Burt (1965) showed that even when the risky shift was produced by discussion without consensus, group members judged the higher risk takers to be more forceful in the group discussion. Once again, however, group members may have reasoned that those who initially favored a risky position must have been more influential.

---

[5] Roger Brown (1965) has completed a thorough discussion of the leadership hypothesis, with particular emphasis on general social-psychological issues.

A more adequate test of the leadership hypothesis was undertaken by Wallach, Kogan, and Burt (1968) who reasoned that if risk takers are generally persuasive, they should be more persuasive than most group members even when discussing issues that have nothing to do with risk taking. Groups, whose members varied widely with respect to their initial level of risk taking, were formed for discussion of issues, some of which contained alternatives that were balanced for riskiness. After the discussions, group members were asked to rank one another for (*a*) degree of influence in the group discussion, (*b*) contribution of the best ideas, and (*c*) greatest provision of effective guidance to the discussions. Among men, high risk takers were not considered more persuasive by any of these criteria, although the relationship was obtained among women.

Evidence from other studies reveals that the leadership hypothesis can hardly qualify as a major explanation of the group-induced shift toward greater risk taking, let alone its sole cause. Wallach, Kogan, and Burt (1967) had groups, homogeneous as to the field dependence or field independence of their members, meet for discussion to consensus of the choice-dilemmas items. The two kinds of groups did not differ with respect to initial risk-taking levels, both field-independent and field-dependent groups shifted toward greater risk taking following discussion to consensus, and the risky shifts were of about the same magnitudes. In addition, high initial risk takers were perceived as more forceful by field-independent subjects but not by field-dependent subjects. The important point for present purposes is that a significant risky shift was found for the latter group, but there was no relationship between initial risk taking level and perceived influence. This result could not be accounted for by the possibility that field-dependent subjects are unable to detect differences in the exertion of influence, for members of field-dependent groups showed significant within-group agreement in their rankings of group members for relative forcefulness in the discussion.

In a related study, Kogan and Wallach (1967d) varied groups in terms of members' test anxiety and defensiveness levels. Despite the fact that the groups shifted toward risk, there was no significant relationship between level of initial risk taking and degree of perceived influence in the discussion. For groups composed of high-defensive, low-anxiety members the reverse was even true: low risk takers were *slightly* more influential than high risk takers. Once again, the result could not be attributed to differential ability of group members to detect degrees of influence.

Brown (1965) reported a study by Nordhøy, who constructed situations on which groups shifted in the opposite direction, that is, toward caution. When the group members were asked to indicate the more persuasive members on these items, they named those who had been more conservative, not more risky, in their initial positions. Rabow, Fowler, Bradford, Hofeller, and Shibuya (1966) found similar results. These findings suggest not that risk takers are more persuasive but that group members attribute greater persuasiveness to those individuals whose initial views were closer to the final position of the group.

A study by Edwards and Willems (personal communication, July 1970) questions the leadership hypothesis. If the risk-prone person is naturally influential in group discussion, his influence should be exerted whether he argues for a risky position or a conservative position.[6] Edwards and Willems asked persons who initially had risky positions to put aside their real views and argue instead for the conservative positions, while persons who were initially conservative were requested to argue for the risky positions. Group members were much more likely to shift toward risk, the position held by initially conservative members who were arguing for the risky positions, than toward caution, the position defended by initially risky members who were arguing for the cautious alternative.

---

[6] Rim (1963, 1964a, 1964b, 1966a, 1966b) suggested that high risk takers are characterized by a pattern of personality traits that predispose them to leadership and influence. Some of the traits of high risk takers are high in extraversion, Machiavellianism, need for achievement, tolerance of ambiguity, radicalism, and interpersonal values of leadership and recognition.

A final study by Vidmar (1970) casts serious doubts on the leadership hypothesis. Experimental groups were formed that were homogeneous as to their initial level of risk taking; that is, all members of a group showed either high, medium, or low risk on the first administration of the questionnaires. If the leadership hypothesis were correct, no shift should have occurred in groups composed of low risk takers because influential high risk persons were absent from the groups. However, a significant shift toward risk was found. In addition, whereas Wallach et al. (1968) found a relation between rated persuasiveness and high risk taking in their female groups, in Vidmar's study female as well as male group members shifted toward risk even when high risk takers were absent.

We can only conclude that the leadership hypothesis is not a major contender among explanations of the risky-shift phenomenon. The finding that high risk takers are perceived to be more persuasive and influential is consistent with the hypothesis. However, as we noted before, caution must be used in attributing causal status to the persuasiveness of high risk takers. The findings (*a*) that group members shift toward greater risk even when high risk takers do not exert greater persuasiveness than low risk takers, (*b*) that it is easier for group members to shift toward risk than toward caution, and (*c*) that risky shifts have been obtained in groups without high risk takers where the leadership hypothesis would predict no shift suggest that other processes than leadership are responsible for the risky-shift phenomenon.

THE DIFFUSION-OF-RESPONSIBILITY HYPOTHESIS

In its present form, the diffusion-of-responsibility hypothesis emphasizes that discussion (*a*) produces emotional bonds between members and (*b*) frees the individual from full responsibility for his later decision because he perceives that his decision has been partially shaped by the group.

The concept of diffusion of responsibility is appealing because of its usefulness in other contexts. It has been used to explain the risky actions of mobs and crowds. The individual's temporary impression of universality among the members of a crowd or mob and his perception of a guaranteed anonymity is used to explain why individuals perform inhuman acts in a group setting where such acts would only bring disgust outside the setting.

Kogan, Wallach, and their collaborators have carried out a number of studies that they feel strongly support the diffusion-of-responsibility hypothesis. In one of these (Wallach et al., 1964), stakes, the probability of achieving these stakes, and group responsibility were varied to reach a clearer understanding of the factors involved in the risky shift. Subjects, who meet in groups of three, were told that their main task would be to answer 10 multiple-choice questions taken from old college board examinations and that they would be paid for correct answers but not for incorrect ones. Furthermore, each subject had to determine beforehand the difficulty level for each of the 10 questions he would answer. The more difficult the question, the more money he would receive for a correct answer.

After the subjects had chosen the level of difficulty for Problems 1–5, the following experimental interventions were introduced for the remaining five questions. In a control condition, subjects merely continued to work individually. In the group-decision condition, subjects were to discuss difficulty levels for the first five problems until they had reached a unanimous group consensus, but each individual would eventually have to solve each problem by himself. In the responsibility-for-group condition, subjects were told that the last five questions would actually be attempted by a randomly determined member of the group, but the winnings of the other two would be entirely dependent upon the outcome of his performance; each subject had to decide on a difficulty level of the five problems before the solver for each was randomly designated, and the solver would receive the problem at the level he had selected. In a group-responsibility–group-decision condition, subjects were to discuss and reach a unanimous decision concerning difficulty levels of the problems to be randomly assigned later. This condition differed from the previous one only in that there was a group discussion. Finally, in a group-designation condition, the group not only determined the difficulty level for each problem but also was to designate a representative who would attempt to solve each problem.

The results showed that the control group differed significantly from each of the other experimental conditions. A significant risky shift occurred in the decision conditions where subjects had to reach a unanimous group decision that would be binding on each member. The designation of an individual to solve a problem had a different effect when it occurred in connection with a group consensus about the problem than when there was no such consensus. In the latter condition, subjects actually became more conservative. Apparently, in the absence of the discussion to consensus, the individual had no way of knowing at what level of difficulty the other individuals would prefer to work; the burden of failure rested squarely on the shoulders of the individual decision maker. Hence, greater conservatism occurred when responsibility for others was created without opportunity for additional communication. However, when the group decided on a particular level of difficulty and then a member was chosen, randomly or by the group, to solve a given problem, there was a maximal diffusion of responsibility. In this case the individual was neither personally responsible for choosing the difficulty level nor responsible for deciding whether to work on a problem. If he was chosen to work on a problem and failed, there would be the consoling fact that the decision had not been his alone.

In a second study, Bem et al. (1965) asked subjects to participate in a study of physiological effects on problem solving. Six experiments were described to three subjects at a time, the subjects' being separated from each other by partitions. The six experiments, involving olfactory stimulation, chromatic stimulation, movement, taste, audition, and sensitivity to odorless gases, were concerned with the effects of various physiological stimulations on the ability to solve simple verbal and mathematical problems. It was explained that subjects would be randomly selected to appear in one of these experiments during their second session. If a subject was selected to participate in one of these six experiments, he could specify the extent to which he was willing to undergo painful temporary side effects so as to obtain more money. The more probable a side effect, the more money he would receive, except that he would be excluded from the study if he actually suffered the side effects. The payments ranged from \$2.80 if the subject chose the alternative that had a 10% probability of producing side effects to \$25 if he chose the alternative with a 90% probability. Subjects choosing the latter probabilities obviously had made risky decisions.

After the subjects filled out the risk-taking questionnaire, the partitions between subjects were removed, and they were requested to discuss each of the six experiments and reach a unanimous decision about the type of stimulation they would endure if the group were chosen to participate in a subsequent experiment as a group. Group decisions were significantly more risky than the mean of the decisions made by the group members as individuals. Furthermore, the private decisions of these same subjects, obtained after completion of a group discussion, had also shifted significantly in a risky direction. Responses in other conditions in the experiment showed that the mere anticipated public disclosure of personal decisions did not lead to a risky shift and that subjects became more conservative when they anticipated undergoing the consequences of the decisions in the company of others or when they anticipated discussion to consensus. The authors concluded that the latter findings clearly support the diffusion-of-responsibility hypothesis at the expense of alternative explanations.

Wallach et al. (1967) systematically composed groups of male adults in terms of field dependence–independence of their members. Although both field-dependent and field-independent groups shifted toward greater risk taking after discussion, longer discussion among field-dependent subjects enhanced the risky shift, while longer discussion among field-independent subjects reduced or even reversed the risky shift. The authors accounted for this result by arguing that field-dependent subjects are more gregarious and affectionate than field-independent subjects. Longer discussion in the former group would emphasize the affective linkage among group members and, hence, the diffusion of responsibility. On the other hand, long discussions by field-independent subjects would emphasize the cognitive analysis of issues, which, according to the authors, would not be conducive to the formation of the emotional ties necessary for diffusion of responsibility to take place. In addition, the authors found a significant positive correlation between the amount of risky shift exhibited by a field-dependent member and the degree of risky shift he attributed to the group of which he was a member. The authors viewed this projection of one's own behavior onto the group as a means of minimizing personal responsibility for one's actions. No such relationship was found for the field-dependent subjects, who presumably are not prone to minimize a sense of personal responsibility for their actions.

Kogan and Wallach (1967d) conducted another personality-oriented study testing the diffusion-of-responsibility hypothesis. They divided female undergraduates into four homogeneous groups: low anxious–low defensive; low anxious–high defensive; high anxious–low defensive; high anxious–high defensive. Significant risky shifts were found for all four groups, but the magnitude of shift varied, with anxious subjects shifting the most and defensive subjects, the least. The authors' interpretation is that homogeneous groups of anxious individuals are uniformly fearful of failure: Discussion of the choice dilemmas makes these individuals more willing to diffuse responsibility in an effort to relieve the burden of possible personal failure. In contrast, homogeneous groups of defensive individuals are quite guarded and careful in their approach to others for fear of exposing personal weaknesses: This orientation interferes with affective interchange among group members and hence tends to inhibit diffusion of responsibility.

The diffusion-of-responsibility hypothesis is strongly supported in these studies by the introduction of personality variables, which presumably reflect the individual's own concern about personal responsibility. In every case, subjects who were expected to shift toward risk when responsibility was diffused did show greater shifts than subjects at the other pole of the relevant personality continuum. Nevertheless, diffusion of responsibility cannot account for all of the variance in the risky-shift studies, for otherwise defensive subjects, field-independent subjects, and subjects with low anxiety would not have shown the significant risky shifts that did occur. In short, the work on personality variables that is supposed to stem directly from the diffusion-of-responsibility hypothesis does not unequivocally support that hypothesis as the exclusive determinant of the risky shift. There is additional evidence that points out weaknesses in the hypothesis and some that contradicts it.

Improving on Wallach and Kogan's (1965) design, Teger and Pruitt (1967) compared information exchange with discussion conditions. In the former condition, subjects gave each other information about their preferred risk levels on various issues without discussing the issues; in the latter, the issues were discussed at some length. The authors found significant risky shifts for both conditions, though the shift was considerably greater in the discussion-to-consensus condition. Kogan and Wallach (1967c) contended that providing subjects with complete information about the views of others is not sufficient to account for the greater risky shifts found in the group discussion condition. However, the results of Teger and Pruitt's study show that discussion is not necessary for a significant shift toward risk to occur. The results of Teger and Pruitt's information-exchange condition have been replicated by Clark and Willems (1969a) and by Willems and Clark (1969, 1971).

Furthermore, the diffusion-of-responsibility hypothesis fails to predict the high correlation that Pruitt and Teger (see Footnote 4) found between the initial riskiness of items and risky shift. Items that initially receive the riskiest judgments produce the most shift. The diffusion-of-responsibility hypothesis would have to predict the opposite, since the low-initial-risk items are mostly those in which the risky alternative has important negative consequences (Clark & Willems, 1969b); thus, initial responsibility should be very strongly felt on these items, and there should be greater effects of diffusion of responsibility.

In addition, Kogan and Wallach have not made explicit the relationship between the development of emotional bonds, formed during discussion, and the negative consequences of the group decision. They stated that discussion allows each member to feel less than proportionately to blame in the event of failure, because he sees others as contributing to the decision with him. How this makes the subject more willing to risk the negative consequences of the decision remains unclear. Does the discussion allow subjects to avoid or repress the negative consequences? Does it somehow lessen the subjective probability of a negative outcome? Or, do the negative consequences actually come to be viewed more positively? Kogan feels that a discussion of the situations lessens the perceived severity of the negative consequences (N. Kogan, personal communication, April 1969). However, empirical research is needed to establish that such is the case.

Willems and Clark (1971) and Pruitt and Teger (1969) have found disconfirming evidence for the relationship between the establishment of emotional bonds among subjects and risky shift. Willems and Clark varied relevancy of the content of discussion orthogonally across conditions of information exchange and discussion. The obtained shifts for subjects who either exchanged risk-relevant information or who discussed risk-relevant content differed significantly from zero; such subjects shifted more toward risk than those who either exchanged irrelevant information or discussed the irrelevant propositions. Since discussion of irrelevant propositions should produce emotional bonds among group members, and since this did not result in a significant shift, Willems and Clark feel that this aspect of the diffusion-of-responsibility hypothesis was refuted.

One could argue that these experimental arrangements did not really test the second part of the diffusion-of-responsibility hypothesis: the implication of others in the final decision. However, a betting experiment conducted by Pruitt and Teger (1969) found no risky shift in a group decision where the members of the group had engaged in past discussions about other issues but were not permitted to discuss or exchange information about the current issue. In this experiment, satisfactory conditions were set for the creation of emotional bonds among group members, and responsibility of the group for the decision was achieved by informing each member that group decisions would be the average of the individuals' decisions. Each individual would have to face the positive or negative outcomes of this averaged decision. No shift was found even when there were emotional bonds among group members and when others were implicated in the group decision. Also, Dion, Miller, and Magnan[7] found in disagreement with the diffusion-of-responsibility hypothesis, that low cohesive groups exhibited greater risky shift than high cohesive groups. At the least, this finding indicates that the creation of emotional bonds is not necessary for the risky shift to occur. The authors, however, do see the finding as consistent with a diffusion-of-responsibility explanation. They suggested that as group members become more attracted to each other, they also become less willing to diffuse personal responsibility onto other group members for the negative consequences of a risky action.

Although the diffusion-of-responsibility hypothesis seems to have been refuted in experiments involving the choice dilemmas or betting, it should be pointed out that it may have explanatory value in other contexts, for instance, the behavior of bystanders in emergencies. Darley and Latané (1968) caused subjects to overhear another subject appear to suffer an epileptic seizure. More subjects from two-person groups responded with help than from three-person groups, and more from three-person than from six-person

[7] Dion, K. L., Miller, N., and Magnan, M. A. Group cohesiveness and social responsibility as determinants of the risky shift. Unpublished manuscript, University of Minnesota, Minneapolis, Minnesota, 1969.

groups. Darley and Latané argued persuasively that the diffusion-of-responsibility hypothesis more appropriately accounts for bystander inaction than the often acclaimed explanations of apathy, alienation, and anomie.[8] When there are several observers present, responsibility for helping is shared among all observers and does not fall to any one person. Also, when others are known to be present, it is convenient to believe that somebody has either already helped the victim or is about to do so. In either case, the result is that no one helps.

Latané and Darley (1968) conducted another experiment on bystander intervention in which the diffusion-of-responsibility hypothesis was less clearly confirmed. Male undergraduates found themselves in a smoke-filled room either alone, with two nonreacting confederates, or in groups of three naive subjects. They found, as predicted, that subjects were less likely to report the smoke when in the presence of passive others or in groups of three than when alone. On the basis of these results and from the postexperimental questionnaire, the authors inferred that the presence of unresponsive bystanders leads the individual to assume that the emergency is not serious and causes him not to react. The authors contend that diffusion of responsibility does not explain these results because they do not see how the "individual's responsibility for saving himself is diffused by the presence of other people [p. 221]." A major difference between this experiment and the previous one was that in the latter case an individual finds himself in danger; in the former, other people seriously need his help. The different results in the two types of situations are intriguing and should be systematically studied.[9]

Notwithstanding the apparent intuitive appeal of the diffusion-of-responsibility hypothesis and its usefulness in accounting for research on bystander intervention, it shows serious shortcomings when applied to the risky shift because: (*a*) Contrary to the hypothesis, discussion is not necessary for the risky shift to occur; (*b*) the hypothesis cannot account for cautious shifts; (*c*) the hypothesis cannot account for the finding that initially risky items are those for which the greatest shift occurs; (*d*) the hypothesis is inconsistent with the finding that the most shift occurs with those items that are perceived to have the least serious consequences; (*e*) the hypothesis does not specify how the creation of emotional bonds among subjects makes them less concerned about the negative consequences of risky decisions; (*f*) finally, most damaging of all, it appears to be the exchange of relevant information, not the development of emotional bonds, that is necessary for the risky shift to occur. In short, it seems rather clear that some other factor in the communication among group members is responsible for producing the shift toward risk.

THE RISK-AS-VALUE HYPOTHESIS

The most persistent and popular interpretation of the risky shift (Brown, 1965; Levinger & Schneider, 1969; Teger & Pruitt, 1967; Wallach & Wing, 1968; Willems, 1969) makes the key assumption that moderate risk is a stronger cultural value than caution and that individuals come to view themselves as being *at least as willing* as their peers to take risks. When a group of such persons meets and discusses matters of risk taking, those whose initial private positions were less risky than the group average will recognize their relative cautiousness and will subsequently recommend greater risk than before in order to restore their perceptions of themselves as relatively risky. On this view, the principal function of the group discussion is to allow group members to compare their positions on the issue to those of other people.

There is considerable evidence that persons perceive themselves to be at least as willing as their peers to take risks. Hinds[10] asked male graduate students in industrial management to try to guess for two choice dilemmas what alternatives would be chosen by other people like themselves. He found that subjects very consistently guessed that others would choose more cautiously than had the subjects themselves. Brown (1965) asked one undergraduate class of 30 students and one graduate seminar of 16 to answer one choice-dilemmas problem for themselves and then to guess how other people like themselves would respond. Not one subject guessed that others would answer more riskily than himself. Brown interpreted this finding as indicating that each individual conceives himself to be at least as risky as the average of his peers.

---

[8] Darley and Latané do not assume a linkage between emotional bonds and diffusion of responsibility; furthermore, the creation of emotional bonds is conceptually unnecessary to account for the behavior of individuals in emergencies.

[9] An excellent summary of the authors' work on bystander intervention in emergencies is provided by Latané and Darley (1969).

[10] Hinds, W. C., Jr. Individual and group decisions in gambling situations. Unpublished master's thesis, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, 1962.

The generality of this finding is affirmed in the following studies. Wallach and Wing (1968), using six of the choice dilemmas, found that both males and females perceived themselves to be more risky than their peers. Levinger and Schneider (1969) found the same results for 9 of the 12 items. Willems (1969) made the same inquiry in the case of a single risk-taking item and found that 81% of his subjects perceived themselves to be at least as willing as their peers to take risks. Willems found the same results in a replication of this experiment. Thus, the data unequivocally reveal a strong and pervasive tendency by persons of both sexes to view themselves as being at least as risky as their peers.

From the risk-as-value hypothesis, however, it would follow that those subjects who perceive themselves as more *cautious* than their peers should not change toward risk following group discussions. Clark, Crockett, and Archer (1971) tested this proposition by comparing discussion groups whose members perceived themselves as being at least as risky as their peers with groups of subjects who perceived themselves to be more cautious than their peers. The results showed that only the former group shifted significantly from zero; the magnitude of shift in the latter group did not differ from that in a control group. Thus, Brown's second assumption was confirmed, for instance, persons who actually shifted toward risk were the ones who perceived themselves to be at least as willing as their peers to take risks.

According to the risk-as-value hypothesis, the main function of the group discussion is informative; it makes clear the level of risk that others will assume in a given situation. Bem et al.'s (1965) experiment on aversive consequences was designed to test this hypothesis. Instead of responding to the choice dilemmas, subjects in a psychophysiological experiment were asked to select the experimental treatment they would endure from a set of treatments that varied in their likelihood to involve disagreeable physical side-effects. The investigators reasoned that in this situation, risk taking would connote the valued properties of courage and boldness and that a risky shift should result when the subjects anticipated that their decisions would be revealed to others. In one condition, subjects were told that each person's decisions would subsequently be made public in the others' presence; in a second condition, subjects were told their decisions would be available to each other so that they could discuss them and make a unanimous selection on just what stimulation they wanted to be employed. The investigators felt that the social desirability of risk taking should be particularly enhanced under these circumstances because the subjects thought they would have to defend their positions in a subsequent discussion. However, no evidence of shift toward risk was found in the first condition, and a conservative rather than a risky shift in the second one. The authors concluded that the shift to risk in group decisions cannot be attributed to a social value on greater risk taking. However, the relevance of this study to the risk-as-value hypothesis must be called into question. As Brown (1965) pointed out, the subjects in both conditions who thought their decisions would be made public were not actually informed of one another's decisions, and since each subject may have already thought himself bold and venturesome, there was no need for him to become more risky.

A more adequate test of the role of information was provided by Teger and Pruitt (1967), who compared groups that discussed the choice dilemmas with groups that merely exchanged information with one another about their initial choices. By the risk-as-value hypothesis, the exchange of information should be enough to show many subjects that they are not as risky as most others in the group; those subjects should then shift toward greater risk. This, in fact, was what Teger and Pruitt found. Similarly, Kogan and Wallach (1967c) found that listening to a tape recording of an actual group discussion led to increased risk taking. In this study, as in the former study, information was made available that permitted the subject, in the absence of discussion, to compare his own risk-taking position with the positions advanced by others. Nevertheless, in both of the studies just cited, simple exchange of information was not as effective in promoting the risky shift as was the group discussion; this suggests that the risk-as-value hypothesis by itself does not account for the phenomenon. However, if one assumes that the discussion condition provides experiences that enhance the value of risk taking, then the centrality of this hypothesis can be defended. The discussion condition not only provides information concerning each individual's position on the issues but it also presents each individual's arguments in defense of his position, making the objectively less risky person more aware of the value he and others place on risk (Teger & Pruitt, 1967).

In support of this interpretation, Lamm (1967) found that subjects who observed a group discussion through a one-way mirror shifted toward greater risk in an amount that did not differ significantly from the shift observed among subjects who actually participated in a group discussion.

Kogan and Wallach's (1967c) listening condition and Lamm's (1967) observation condition might also be interpreted as supporting the diffusion-of-responsibility hypothesis since the groups did, after all, engage in discussion. It is possible to untangle this theoretical issue by initiating experiments in which one of the hypotheses makes a prediction that the other one does not. Vidmar (1970) has done such a study. Three relatively homogeneous sets of groups, consisting of high, medium, and low risk takers, were compared with mixed groups composed of two high, one medium, and two low risk takers. All four types of groups showed the risky shift, but the mixed, or heterogeneous groups showed significantly greater shifts than the more homogeneous groups; the three homogeneous groups did not differ significantly from one another in magnitude of risky shift. Vidmar also found that the magnitude of individual risky shift was inversely related to the member's initial risk position in the group. That is, the objectively conservative members in a group shifted more than did the risky members, exactly as the risk-as-value hypothesis predicts. It should be pointed out that the familiarization hypothesis and the diffusion-of-responsibility hypothesis cannot account for the twin facts that the heterogeneous groups shifted more toward risk than the homogeneous groups and that the objectively conservative subjects shifted more than the objectively risky members.

Willems and Clark (1971) varied homogeneous and heterogenous groups of females orthogonally across conditions of information exchange and discussion, using the six choice-dilemmas items that had produced the most risky shift in previous studies. The criteria for group compositions were as follows: On three critical items (4, 6, and 7 by Kogan and Wallach's numbering) each person in the homogeneous groups did not differ by more than two scale points. For heterogeneous groups, at least two persons differed by at least four scale points on each of the three items. Under both information-exchange and discussion conditions, only the heterogeneous groups gave shifts significantly greater than zero. These findings were interpreted as reflecting the fact that in heterogeneous groups the probability was increased that some subjects would discover that their choices were not as risky as their peers, producing motivation to shift. In the homogeneous groups, in which subjects differed by no more than two scale points, there was no real need for persons to shift toward risk because nobody was less risky than his peers; all subjects were already actualizing the value of risk.

The results of the study by Willems and Clark differ in one important way from those reported by Vidmar. While both found significant shifts in heterogeneous groups, Willems and Clark did not find a significant shift toward risk in the homogeneous groups, as

Vidmar did. The difference may be due to a subtle difference in procedure. Vidmar formed his groups of high, medium, and low risk takers by summing each subject's scores over 10 choice dilemmas. Summing scores over items, however, can conceal diversity of responses on certain key items, like Items 4 and 7 of the Wallach and Kogan questionnaire. Since these items seem to be responsible for most of the shift found in a variety of studies, Willems and Clark argued that persons in homogeneous groups must respond similarly on those key items. Otherwise, apparently homogeneous groups may, in fact, be heterogeneous.

There is empirical evidence that supports this contention. Using the same criteria as Vidmar (1970), Hoyt and Stoner (1968) found significant shifts toward risk in their homogeneous groups, while the magnitude of shift in these groups did not differ significantly from that in previous studies in which groups were randomly composed and, therefore, assumed to be heterogeneous. At the same time, Hoyt and Stoner found a significant positive relation between dispersion of individual preferences on an item and the magnitude of group shift. They remarked:

> Thus it appears that homogeneous grouping reduced the ranges of total individual risk scores within each group very substantially (by almost three-fourths) but did not reduce individual item ranges very appreciably (by about one-seventh) [Hoyt & Stoner, 1968, p. 281].

Pruitt and Teger (see Footnote 4) cited other evidence supporting the risk-as-value hypothesis. They compared initial risk level with amount of risky shift across the items in a number of studies and found a very high correlation between the initial position on an item and the risky shift for that item. That is, judgments on items that elicit a risky response initially become even more risky after discussion; items that elicit a cautious response initially become even more cautious after discussion. Pruitt and Teger argued that this finding favors the risk-as-value hypothesis, on the proposition that the average initial position on an item indicates whether the value of risk or the value of caution is appropriate to that issue. Clark and Willems (1969b) also found a very high correlation between the mean initial responses to the 12 items of the Kogan and Wallach questionnaire and the national norms of risky shift for these items.

Further compelling evidence for the risk-as-value hypothesis comes from a betting experiment by Blank (1968). Subjects chose bets in either a group or an individual situation. Under individual conditions, subjects chose a significantly greater number of high-probability bets than did subjects in the group condition. More importantly, subjects who transferred from the group condition to the individual condition did not differ significantly in the riskiness of their bets from subjects who continued in the individual condition; however, subjects who transferred from the individual condition to the group condition showed a significant shift to greater risk. This study clearly indicated that the risky shift occurs in games of chance, that it may occur even when there has been no group discussion, and that it does not endure when subjects are switched to an individual condition.[11] Since no discussion was allowed in the groups, Blank asserted that the diffusion-of-responsibility hypothesis could not easily explain this finding.

The most attractive feature of the risk-as-value hypothesis is that it can be easily extended to account for items on which judgments consistently shift toward either risk or caution and for items on which no consistent shift in judgments is evident. If an issue invokes values favoring the risky alternative, then the hypothesis would predict that most individuals would consider themselves to be at least as willing as their peers to take risks, so that the effect of group discussion would be to produce a shift toward risk. For those items that elicit cautious values, the hypothesis predicts that individuals will perceive themselves to be at least as cautious as their peers, and we would expect a shift toward caution as a result of the group discussion. Stoner (1968) has performed a study that confirms these predictions. He presented his subjects with a 12-item questionnaire in which 6 items were assumed to elicit risky values and 6 to elicit cautious values. Subjects also completed a value-rating instrument that required them to rank in order of importance 18 phrases that reflected values that were implicit in the outcomes of the 12 dilemmas items. On the basis of the subjects' rankings, the 12 dilemmas items were divided into two groups: those for which the risky alternatives implied values that were ranked higher in importance than the cautious alternatives and those for which the opposite was true.[12] Subsequently, subjects discussed items in groups and arrived at a group decision. Results generally supported the risk-as-value predictions. Subjects were more risky initially on risk-oriented items and more cautious initially on cautious-oriented items. Significant differences between individuals' perceptions of their own and others' riskiness were also found. On risk-oriented items, subjects judged themselves to be relatively risky and, after unanimous group decisions, gave responses that were more risky than the average of the initial individual decisions. On caution-oriented items, however, the results were less clear-cut. Subjects perceived themselves to be more cautious than most people on only three of six cautious-oriented items; of the remaining items, two showed nonsignificant judgments in the direction of caution, while the third showed nonsignificant judgments in the direction of risk. In addition, only two of the six cautious-oriented items resulted in cautious shifts, the other four showing nonsignificant changes. Whether these results reflect the nature of the specific items that were employed or the fact that risk is a stronger value than caution in our culture is a matter for future research.

---

[11] Kogan and Wallach (1967a) reported several studies using the choice dilemmas where the shift did endure after discussion.

Stoner's assessment of the importance of values related to the risky and cautious items was an attempt to solve one of the major problems with the risk-as-value hypothesis, that of distinguishing in advance those issues on which a shift to risk will occur from those associated with a shift to caution. An additional distinction can be made between risk-oriented and caution-oriented items. Clark and Willems (1969b) have suggested that the perceived consequences of failure in each situation play a crucial role in determining the level of risk subjects will take in that situation. They asked subjects to assume that the hypothetical person in each situation chose the attractive alternative and that it failed. Subjects were asked to indicate how serious the consequences would be. The mean ratings of the consequences of failures were then correlated with Pruitt and Teger's national norms for the likelihood of risky shifts; the two sets of ratings correlated very highly ($r = -.70$, $p < .01$); that is, the greatest risky shifts occur for items with the least serious perceived consequences. In addition, the mean riskiness of subjects' initial positions on the items was correlated with the perceived consequences: those items that initially elicit risk are perceived to have relatively trivial consequences, those that initially elicit caution are perceived to have severe consequences. At the least, these data suggest that subjects' perceived consequences of failure play a part in their proneness toward risk and caution.

Taken together, the evidence available converges to suggest that (*a*) treatment of risk-relevant content, (*b*) in situations for which the consequences of failure are not severe, (*c*) in groups whose members vary in their initial riskiness, and (*d*) for individuals who perceive themselves to be at least as willing as their peers to take risks *jointly* represent the necessary and sufficient conditions for the risky shift. This is so because when all four of these conditions are not met, the shift occurs (sufficient conditions) and when *any* of the four conditions is not met, the shift does not occur (necessary conditions). Since three of the conditions (except Condition *b*) are direct derivatives of the risk-as-value hypothesis, the evidence is strong support for that view.[13]

---

[12] If risk-taking items do involve values and if the importance of risky and cautious alternatives can be determined, then perhaps we may have a way of distinguishing between those items that produce cautious shifts and those items that produce risky shifts—hence, the risk-as-value hypothesis would not be an ex post facto explanation since subjects' perception of their own relative riskiness would tend to be consistent with and a good predictor of the subsequent group shift toward either caution or risk.

[13] Jellison and Riskind (1970) have recently completed studies that relate risk-as-value theory to Festinger's theory of social comparison.

Current Status of the Risky-Shift Phenomenon

The empirical shift toward risk using the choice dilemmas as a measure can be shown to have at least four main characteristics: (*a*) Regular and systematic average shift toward risk; (*b*) the convergence of risk preferences; (*c*) the ceiling effect, that is, that no individual's risk preference exceeds the preference of the riskiest individual before the group process occurs;[14] and (*d*) the preponderance of comments in the discussion favors risk (Brown, 1965).

On a post hoc basis, the familiarization hypothesis would seem to account for the systematic shift but none of the other three aspects. Not only does the familiarization hypothesis fail to account for the convergence of views, the ceiling effect, and the preponderance of comments favoring risk but, given the hypothesis, one cannot derive predictions concerning these key characteristics of the risky shift. This is because these characteristics refer to group processes, and the familiarization hypothesis regards the risky-shift phenomenon as a pseudogroup phenomenon. In addition, recent studies show that the original familiarization findings cannot be replicated.

On a post hoc basis, the leadership hypothesis, with its emphasis on the leader's persuasibility, would appear to account for the systematic shift, the ceiling effect, the convergence around the opinion of the leader, and the preponderance of comments favoring risk. However, on an empirical basis, the review of studies has shown that derivations from the leadership hypothesis do not predict and account for the phenomenon.

On a post hoc basis, the diffusion-of-responsibility hypothesis would seem to account for the systematic shift and perhaps the preponderance of comments in the discussion favoring risk, but not the convergence of views or the ceiling effect. The former two characteristics can be derived from the hypothesis but not the latter two, since it is possible for each person in a group to become more risky without changes in variability occurring as a result. Since the hypothesis emphasized generalized group processes, a prediction concerning the ceiling effect cannot be made. In fact, one might wonder why, as a result of diffusion of responsibility, the final decisions are not riskier than the riskiest individual. Although the hypothesis does predict the systematic shift toward risk and the preponderance of comments favoring risk, experimental studies reported above, especially the ones involving relevance of content, perception of self as relatively risky, and diversity of opinion, suggest that diffusion of responsibility as a postulated process is not necessary or sufficient to produce the shift.

In contrast to the other three hypotheses, on a post hoc basis, the risk-as-value hypothesis would account for all four aspects of the risky shift. The hypothesis predicts: (*a*) The systematic shift toward risk occurs when persons are not, in fact, expressing positions at least as risky as the norm; (*b*) the convergence of risk preferences around a central value is the result of value attainment; (*c*) the reason for the ceiling effect is that persons want to be at least as risky as their peers, and the fact that most individuals move toward the riskiest individual achieves this; and (*d*) the preponderance of discussion favors risk because the discussion of risk-relevant material allows the individual to compare his level of risk taking with that of others. Furthermore, the experimental studies reported above, particularly the ones in relevance of content, heterogeneity of initial opinions, and perception of riskiness, demonstrate that risky shifts occur under the conditions specifically predicted by the hypothesis. Finally, this hypothesis has the potential advantage of being able to account for cautious as well as for risky shifts.

However, one troublesome problem is the dependence of the risky shift on instructions. The stability and replicability of the risky-shift phenomenon have been firmly established. In fact, most investigators begin by assuming its stability and have been searching for explanations of the shift. The two most popular interpretations, the diffusion-of-responsibility and risk-as-value hypotheses, both point to group processes as producing the risky shift. That is, as they now stand, both interpretations are stated in terms that are independent of the specific experimental instructions used to elicit risk preferences. However, there is evidence from recent experiments by Clark and Willems (1969a) and Willems and Clark (1969) that the risky shift is a function of the instructions that are normally used in experiments on this topic.

---

[14] In the author's research, among 1,320 events (220 Subjects × 6 Choice-Dilemmas Items) only 57 outcomes involved final decisions that were more risky than the initial positions of the riskiest members of the groups. In other words, only 4% of the possible outcomes violated this assertion of a ceiling effect.

Risk taking was measured by the Kogan and Wallach choice-dilemmas instrument, the format most widely used in the research on the risky shift. On this questionnaire, the experimental subject's task is to serve as a hypothetical adviser to the focal person by choosing the *lowest probability* of success that he would consider acceptable for the hypothetical person to select the more attractive risky alternative. Clark and Willems reasoned that the phrase *lowest probability* in the instructions may orient either the individual in his responding or the group in its discussion toward riskiness and that more neutral instructions might reduce the risk-shift effect. Half of the subjects were given neutral instructions that requested them simply to choose, for each item, the probability of success they would consider acceptable to recommend the more attractive alternative; the other half of the subjects received the standard instructions. These two sets of instructions, risk-oriented and neutral, were varied orthogonally across information-exchange, discussion, listening, and control conditions. The results showed a significant risky shift only for the groups that received the standard risky instructions. Since no significant shifts occurred in any of the groups that received the neutral instructions, the authors concluded that the risk-oriented instructions provide directional expectations for risk preferences.[15] One might still argue, however, that since risky shifts have occurred in varying magnitudes from different manipulations, other concepts have to be formulated for a complete understanding of the risky-shift phenomenon.

Furthermore, the findings on instruction dependence do nothing to distinguish between the two major hypotheses now used to account for the risky shift. Diffusion-of-responsibility proponents might suggest that the lowest probability instruction becomes salient in the discussion and enhances the feeling of less potential blame. The risk-as-value proponents would probably suggest that the instructions operate to persuade subjects to move in a direction of the more risky alternatives. However, the findings do suggest that some conceptual and methodological restrictions must be placed on the range of generalizations we can make about the phenomenon. Specifically, the findings suggest that care must be exercised in interpreting the Kogan and Wallach questionnaire as a measure of risk taking. Until further research is conducted on instruction dependence and risk choices involving other experimental formats—for example, betting with payoffs—we cannot generalize instruction dependence to them. The importance of such research should be clear.

These data suggest, then, that instruction dependence must be added to relevance of context, in situations for which the consequences of failure are not severe, heterogeneity of opinions and distribution of perception of peers as a condition of the risky shift. When those conditions *and* risky-oriented instructions occur together, the risky shift will be obtained; when any of the four conditions is not met, it seems, the shift will not occur. Until recently, this seemed to be the case. But, as noted below, new research has raised some new and troublesome issues.

One difficulty, as discussed earlier, with the risk-as-value hypothesis as it was originally formulated by Brown is that it cannot account for the discrepancies found between information-exchange, listening, and discussion conditions. Teger and Pruitt (1967), however, argued for a modification of Brown's hypothesis, pointing out that in the discussion condition there is greater opportunity for group members to acquaint themselves with the arguments in favor of the value of risk.

---

[15] The finding that risk-oriented instructions generate conservative rather than risky shifts, or do not cause shifts, for materials of certain kinds, is not a serious contradiction to an instructional demand characteristic interpretation of the risky shift, but it does suggest that instructions are not the only factor that generate shifts. Wallach and Mabli's (1970) view that Clark and Willems' neutral instructions encourage a wider range of choices and induce a force toward compromise and concession ignores the finding that there were no differences between the risk-oriented and neutral-oriented instructions in initial responses or variability.

Clark et al. (1971) performed an experiment relevant to this issue. They had subjects either (*a*) exchange information about their choices on the risk dilemmas; (*b*) discuss the dilemmas and indicate their choices on each item and their arguments for those choices; or (*c*) discuss the choice dilemmas, but only give their arguments in favor of the various alternatives without specifying which alternatives they had chosen. In both discussion conditions, judgments shifted significantly toward risk, in amounts significantly greater than in the information-exchange condition, which did not differ significantly from zero. This result appears to be inconsistent with the risk-as-value hypothesis, for since risk is a value, and we perceive ourselves to be at least as willing as our peers to take risks, we should need to reevaluate our own positions only when we actually discover that others choose riskier alternatives than we do; however, a significant shift occurred in the discussion condition when numerical preferences for specific alternatives were not exchanged. The hypothesis can be rescued only by assuming that listening to and participating in a discussion of arguments in favor of risk, even without clear commitment from each subject to an alternative, permit the discussant to learn that his peers are riskier than he had thought. There is evidence from this previous study, however, that suggests that subjects did change their perceptions of others' responses to specific alternatives. Subjects indicated before and after discussion of the choice dilemmas what they thought the majority of people like themselves would mark for each item. Subjects in both discussion groups—with and without exchange of the alternatives chosen—came to change their judgments to represent other people as riskier than initially; the two conditions did not differ significantly. Thus, the data do suggest that the mediation of the cultural value of risk through the specification of probabilities is still tenable.[16] However, a more adequate test is needed to substantiate this view.

Nevertheless, the idea that information exchange leads to social comparison processes is still a troublesome possibility. According to the risk-as-value hypothesis, the risky shift occurs because the individual discovers that he is not as risky as his peers; in order to maintain the implicit norm of riskiness, he becomes more risky. Recent evidence, however, suggests that risky and cautious shifts do not occur when social comparison processes and elicitation of arguments in favor of risky or cautious items, respectively, are not biased toward the implicit values of the decision items. In a pilot investigation, Steiner (cited in Dion, Baron, & Miller, 1970), using both risk-oriented and cautious-oriented items, presented false feedback that disconfirmed the subjects initial expectation concerning the responses of others. The manipulated norms produced marked shifts of own riskiness in the direction of the position advocated in the false feedback.

Baron, Dion, and Baron[17] conducted a similar study. Subjects discussed either risk-oriented or caution-oriented items. Using confederates, the investigators created a majority consensus that was either two scale positions riskier or more conservative than the naive subject's initial decision. Subjects exhibited marked conformity to the artificially established group consensus, regardless of whether the group norm coincided or conflicted with the underlying value elicited by the type of item. Furthermore, these shifts remained intact in private, posttest measures of risk taking.

Clark and Crockett (1971) had low, medium, or high risk takers listen to contrived tape recordings of discussion groups composed of low, medium, or high risk takers on six risky items. The results were striking.

---

[16] Madaras and Bem (1968) presented evidence that they interpreted as indicating that cultural value of risk is mediated by the knowledge of others' arguments, for instance, information about risk alternatives.

[17] Baron, R. S., Dion, K. L., and Baron, P. Group norms, elicited values, and risk-taking. Unpublished manuscript, University of Minnesota, Minneapolis, Minnesota, 1968.

Subjects always conformed to the position advocated by the taped discussion groups, regardless of whether the positions advocated coincided with the underlying value of risk.

These results do not argue well for an interpretation that assumes that social comparison pressures operate in conjunction with elicited values. If risk and caution are strong cultural values, individuals should selectively adjust their levels of risk and caution to coincide with the dominant values underlying risk-oriented and caution-oriented items. Yet, the previous studies showed that subjects conform dramatically to the majority consensus provided them. Thus, it appears that the social comparison mechanism of value theory fares poorly.

Evidence from a recent experiment makes this conclusion premature. Wallach and Mabli (1970), testing the social comparison mechanism process of value theory against conformity, composed three-person groups that contained either a conservative minority and risky majority or a conservative majority and risky minority. As predicted by value theory, in contrast to the conformity interpretation, whether they constituted a minority or a majority of the group membership, conservatives showed strong and similar shifts toward greater risk as a result of discussion of the choice dilemmas. In contrast, whether they constituted a minority or a majority of the group membership, risk takers showed essentially no shift as a result of discussion. For the results of Steiner, of Baron, Dion, and Baron and of Clark and Crockett to hold up, one would have expected conservatives not to shift toward risk when they were in a majority, and risky persons to shift toward caution when they were in the minority. Procedural differences in how the groups were constituted may account for the discrepancies in results. Wallach and Mabli did not directly manipulate scale position in their groups, particularly on an item-by-item basis as did the other investigators. In addition, Wallach and Mabli's risky subjects had the opportunity to persuade the more cautious subjects.

We can only conclude that individuals are culturally disposed to generate and favor risky arguments when considering risky items and cautious arguments when considering cautious items (Brown, 1965; Madaras & Bem, 1968), but when forced to consider cautious arguments and judgments on risky items and risky arguments and judgments on cautious items, individuals do change their judgments in the direction advocated by the arguments to minimize the discrepancy between themselves and others. Whether these results hold up when an individual is exposed to more heterogeneous judgments and arguments (persons in the above experiments heard only judgments and arguments that were homogeneously discrepant from their own), selected so that the individual is exposed to views that are the same or slightly different from his own as well as those that are very different, will have to wait for further experimental analysis. Such further work is important because, as they stand, these results raise questions about the risk-as-value hypothesis and the risky-shift phenomenon itself.

Future research is needed to determine whether modifications of the above five necessary conditions are needed or if additional conditions have to be found to account for situations where social comparison pressures and elicitation of arguments that are not biased toward the implicit values of the risk-taking dilemmas. At present, it is particularly unclear just what distribution of arguments (percentage of risky and cautious arguments) must occur for risky or cautious shifts to result.

REFERENCES

Allport, F. H. *Social psychology*. Boston: Houghton, Mifflin, 1924.

Asch, S. E. Forming impressions of personality. *Journal of Abnormal and Social Psychology*, 1946, **41**, 258–290.

Asch, S. E. Effects of group pressure upon the modification and distortion of judgments. In E. E. Maccoby, R. M. Newcomb, & E. L. Hartley (Eds.), *Readings in social psychology*. New York: Holt, Rinehart & Winston, 1958.

Barnlund, D. C. A comparative study of individual, majority, and group judgment. *Journal of Abnormal and Social Psychology*, 1959, **58**, 55–60.

Bateson, N. Familiarization, group discussion, and risk taking. *Journal of Experimental Social Psychology*, 1966, **2**, 119–129.

Bem, D. J., Wallach, M. A., & Kogan, N. Group decision making under risk of aversive consequences. *Journal of Personality and Social Psychology,* 1965, **1**, 453–460.

Blank, A. D. Effects of group and individual conditions on choice behavior. *Journal of Personality and Social Psychology,* 1968, **8**, 294–298.

Brown, R. *Social psychology.* New York: The Free Press, 1965.

Cartwright, D., & Zander, A. (Eds.) *Group dynamics.* New York: Harper and Row, 1960.

Clark, R. D., & Crockett, W. H. Subjects initial positions, exposure to varying opinions, and the risky shift. *Psychonomic Science,* 1971, **23**, 277–279.

Clark, R. D., III, Crockett, W. H., & Archer, R. L. The relationship between perception of self, others, and the risky shift. *Journal of Personality and Social Psychology,* 1971, in press.

Clark, R. D., III, & Willems, E. P. Where is the risky shift? Dependence on instructions. *Journal of Personality and Social Psychology,* 1969, **13**, 215–221. (a)

Clark, R. D., III, & Willems, E. P. Risk preferences as related to judged consequences of failure. *Psychological Reports,* 1969, **25**, 827–830. (b)

Collins, B. E., & Guetzkow, H. A. *A social psychology of group processes for decision making.* New York: Wiley, 1964.

Darley, J. M., & Latané, B. Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology,* 1968, **8**, 377–383.

Dion, D. L., Baron, R. S., & Miller, N. Why do groups make riskier decisions than individuals. In L. Berkowitz (Ed.), *Advances in experimental social psychology.* Vol. 5. New York: Academic Press, 1970.

Farnsworth, P. R., & Behner, A. A note on the attitude of social conformity. *Journal of Social Psychology,* 1931, **2**, 126–128.

Festinger, L. A theory of social comparison processes. *Human Relations,* 1954, **7**, 117–140.

Flanders, J. P., & Thistlethwaite, D. L. Effects of familiarization and group discussion upon risk taking. *Journal of Personality and Social Psychology,* 1967, **5**, 91–97.

Hoyt, G. C., & Stoner, J. A. F. Leadership and group decisions involving risk. *Journal of Experimental Social Psychology,* 1968, **4**, 275–284.

Jellison, J. M., & Riskind, J. A social comparison of abilities interpretation of risk-taking behavior. *Journal of Personality and Social Psychology,* 1970, **15**, 375–390.

Jones, E. E., & Gerard, H. B. *Foundations of social psychology.* New York: Wiley, 1967.

Kelley, H. H., & Thibaut, J. W. Group problem solving. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology.* Vol. IV. Reading, Mass.: Addison-Wesley, 1969.

Kogan, N., & Wallach, M. A. *Risk taking: A study in cognition and personality.* New York: Holt, Rinehart & Winston, 1964.

Kogan, N., & Wallach, M. A. Risk taking as a function of the situation, the person, and the group. In, *New directions in psychology.* Vol. III. New York: Holt, Rinehart & Winston, 1967. (a)

Kogan, N., & Wallach, M. A. Effects of physical separation of group members upon group risk taking. *Human Relations,* 1967, **20**, 41–48. (b)

Kogan, N., & Wallach, M. A. The risky-shift phenomenon in small decision-making groups: A test of the information-exchange hypothesis. *Journal of Experimental Social Psychology,* 1967, **3**, 75–84. (c)

Kogan, N., & Wallach, M. A. Group risk taking as a function of members' anxiety and defensive levels. *Journal of Personality,* 1967, **35**, 50–63. (d)

Krech, D., Crutchfield, R. S., & Ballachry, E. L. *Individual in society: A textbook of social psychology.* New York: McGraw-Hill, 1962.

Lamm, H. Will an observer advise higher risk taking after hearing a discussion of the decision problem? *Journal of Personality and Social Psychology,* 1967, **6**, 467–471.

Latané, B., & Darley, J. M. Group inhibition of bystander intervention in emergencies. *Journal of Personality and Social Psychology,* 1968, **10**, 215–221.

Latané, B., & Darley, J. M. Bystander apathy. *American Scientist,* 1969, **2**, 244–268.

Levinger, G., & Schneider, D. J. A test of the risk is a value hypothesis. *Journal of Personality and Social Psychology,* 1969, **11**, 165–169.

Madaras, G. R., & Bem, D. J. Risk and conservatism in group decision-making. *Journal of Experimental Social Psychology,* 1968, **4**, 350–365.

Marquis, D. G. Individual responsibility and group decisions involving risk. *Industrial Management Review,* 1962, **3**, 8–23.

Pruitt, D. G., & Teger, A. I. The risky shift in group betting. *Journal of Experimental Social Psychology,* 1969, **5**, 115–126.

Rabow, J., Fowler, F. J., Bradford, D. L., Hofeller, M. A., & Shibuya, Y. The role of social norms and leadership in risk taking. *Sociometry,* 1966, **29**, 16–27.

Rim, Y. Risk taking and need for achievement. *Acta Psychologica,* 1963, **21**, 108–115.

Rim, Y. Personality and group decision involving risk. *Psychological Record,* 1964, **14**, 37–45. (a)

Rim, Y. Social attitudes and risk taking. *Human Relations,* 1964, **17**, 259–265. (b)

Rim, Y. Who are the risk-takers in decision making? *Personnel Administration,* 1966, **29**, 26–30. (a)

Rim, Y. Machiavellianism and decisions involving risk. *British Journal of Social and Clinical Psychology,* 1966, **5**, 30–36. (b)

Schachter, S. Deviation, rejection, and communication. *Journal of Abnormal and Social Psychology,* 1951, **46**, 190–207.

Schein, E. H. *Organization psychology.* Englewood Cliffs, N. J.: Prentice-Hall, 1965.

Secord, P. F., & Backman, C. W. *Social psychology.* New York: McGraw-Hill, 1964.

Sherif, M. A study of some social factors in perception. *Archives of Psychology,* 1935, No. 187.

Sherif, M. *The psychology of social norms.* New York: Harper, 1936.

Stoner, J. A. F. Risky and cautious shifts in group decisions: The influence of widely held values. *Journal of Experimental Social Psychology,* 1968, **4**, 442–459.

Teger, A. I., & Pruitt, D. G. Components of group risk taking. *Journal of Experimental Social Psychology,* 1967, **3**, 189–205.

Teger, A. I., Pruitt, D. G., St. Jean, R., & Haaland, G. A. A reexamination of the familiarization hypothesis in group risk taking. *Journal of Experimental Social Psychology,* 1970, **6**, 346–350.

Vidmar, N. Group composition and the risky shift. *Journal of Experimental Psychology,* 1970, **6**, 153–166.

Wallach, M. A., Kogan, N., & Bem, D. J. Group influence on individual risk taking. *Journal of Abnormal and Social Psychology,* 1962, **65**, 75–86.

Wallach, M. A., Kogan, N., & Bem, D. J. Diffusion of responsibility and level of risk taking in groups. *Journal of Abnormal and Social Psychology,* 1964, **68**, 263–274.

Wallach, M. A., & Kogan, N. The roles of information, discussion, and consensus in group risk taking. *Journal of Experimental Social Psychology,* 1965, **1**, 1–19.

Wallach, M. A., Kogan, N., & Burt, R. B. Can group members recognize the effects of group discussion upon risk taking? *Journal of Experimental Social Psychology,* 1965, **1**, 379–395.

Wallach, M. A., Kogan, N., & Burt, R. B. Group risk taking and field dependence-independence of group members. *Sociometry,* 1967, **30**, 323–338.

Wallach, M. A., Kogan, N., & Burt, R. B. Are risk takers more persuasive than conservatives in group discussion? *Journal of Experimental Social Psychology,* 1968, **4**, 76–88.

Wallach, M. A., & Mabli, J. Information versus conformity in the effects of group discussion on risk taking. *Journal of Personality and Social Psychology,* 1970, **14**, 149–156.

Wallach, M. A., & Wing, C. W., Jr. Is risk a value? *Journal of Personality and Social Psychology,* 1968, **9**, 101–106.

Willems, E. P. Risk is a value. *Psychological Reports,* 1969, **24**, 81–82.

Willems, E. P., & Clark, R. D., III. Dependency of risky shift on instructions: A replication. *Psychological Reports,* 1969, **25**, 811–814.

Willems, E. P., & Clark, R. D., III. Shift toward risk and heterogeneity of groups. *Journal of Experimental Social Psychology,* 1971, **7**, 304–312.

Zander, A., & Medow, H. Individual and group levels of aspiration. *Human Relations,* 1963, **16**, 89–105.

(Received December 15, 1970)

Psychological Bulletin
1971, Vol. 76, No. 4, 271–294

PERSONALITY AND ORGANIZATION THEORY:
A REVIEW OF SOME CONCEPTUAL LITERATURE

CARY M. LICHTMAN[1]
Wayne State University

AND

RAYMOND G. HUNT
State University of New York at Buffalo

Since a theory of organization can be no better than the assumption it makes about the human personality, this paper reviews and comments on the ways different theories have dealt with, or contributed to, the "structuralistic" versus "personalistic" dilemma. The several views are classed under four conceptual rubrics: (*a*) traditional structural approaches (e.g., Marx and Weber), (*b*) modern structural approaches (e.g., Argyris and McGregor), (*c*) personalistic views (e.g., perceptual theories), and (*d*) integrating approaches (e.g., system and role models). It is concluded that the extreme variability found within and among organizations renders the one-sided normative theories less useful in understanding organizational behavior than models that recognize situational contingencies.

When viewing the phenomena of human behavior, it has been a custom among social scientists to debate the relative importance to it of individual characteristics and properties of the social situation. Sometimes they hide from the problem, and, of course, accommodations have been sought, but the question still divides "personalists" and "situationalists" despite the efforts of a growing school of "interactionist" peacemakers (cf. Hollander & Hunt, 1967, especially Sections I and III for fuller discussion of these issues).

Theories of organization, dealing as they do with human collective undertakings, could hardly expect to escape the conceptual fallout from this "debate." Nor have they. And because we believe a theory of organization can be no better than the assumptions it makes about the human personality, we propose to review and comment on the ways different theories have chosen to resolve (or evade) the issue of relating persons and organizations.

In this discussion we construe organizations, after W. R. Scott (1964), as "collectivities . . . established for the pursuit of relatively specific objectives on a more or less continuous basis [p. 488]." Although organizational policies and structures may be elaborated separately from their memberships, people populate and work them, and the interplay of people and structures defines the phenomena of organization (cf. Hollander & Hunt, 1967, pp. 352–353).

Early theories of social organization emphasized the more general sociological conviction that the social structure was the primary determinant of differential human characteristics. Classical organization theorists, therefore, proposed that these structural determinants be utilized in the design of organizations to maximize human potential and organizational efficiency. Some two decades later, however, the so-called neoclassical organization theorists offered models based on the personalistic views of psychology. Reacting to early sociological notions, this school of thought rejected the formal organizational structure as of small importance and placed the locus of the organization within the phenomenal fields of its individual incumbents. Then, reacting in turn to the one-sidedness of these two earlier approaches, many modern organization theorists acknowledged the value of each and attempted to integrate them into a unitary systemic conceptual scheme. Thus, modern organization theory (cf. Scott, 1961) proposes that human behavior in organizational settings can be understood in terms of three elements: (*a*) the stated design of functions, that is, the requirements of the organization; (*b*) the characteristics of people who populate the organization, that is, the attributes they bring with them into the organization, including those derived from other social affiliations; (*c*) the relations between the organization's defined properties and the characteristics of people who populate it. The last element is a system concept that provides for prospects of organization-level emergents and is a hallmark of "modern" organization theory.

---

[1] Requests for reprints should be sent to C. M. Lichtman, Department of Psychology, Wayne State University, Detroit, Michigan 48202, or R. G. Hunt, Department of Psychology, State University of New York, Buffalo, New York 14226.

In their entirety, particular theories of organization do not always fit neatly into any one of these categories. The limited conceptual scope of some theories obviously places them with either the structural or personalistic camps; others seem to be a combination of both, without necessarily being full-fledged systemic theories. Grouping conceptual formulations into categories may not always do maximum justice to many of their nuances, but it does help to integrate the diverse ideas contained in the organizational literature. And the task of grouping theories becomes somewhat simpler if the classification is done, as we have tried to do here, according to the actual content of particular theoretical statements rather than what their authors claim to be the domain of their conceptualizations.

VARIETIES OF THEORY

For classificatory purposes we elected to employ four categories instead of the traditional three. The first two sections of this discussion deal in sequence with *traditional structural approaches* and *modern structural approaches* to organizational behavior. What the old and new structural theories have in common is the use of a global theory of personality as a premise upon which an organizational structure is designed. "All men are thus," they say, "and organizational theory and practice had best take account of such truths [Quinn & Kahn, 1967, p. 451]." In the case of traditional structuralists (classical organization theorists) the grounding assumption is that man is lazy, untrustworthy, and works only for money. The resulting organizational design suggested by this assumption, of course, has been of the bureaucratic variety, and the motivational presuppositions of traditional structuralism still underly much modern "incentive" theory and associated modes of compensating performance (cf. Opsahl & Dunnette, 1966).

Modern structural theorists follow quite the same logic, although basing it on a different initial premise. They argue that all men are interested in self-actualizing or realizing their full potential, and to allow for this they recommend a looser, more decentralized structure than did the bureaucrats. Both views are logically committed to the position that since people share certain important characteristics, differences in people or performance can best be explained in terms of differential positional occupancy in the organizational structure. The implication of this typology is that theorists such as Argyris and McGregor (and some others as well), who have usually been considered personalistic theorists, emerge as structural theorists in the present context.

The third section of the discussion deals with those theories that emphasize the *personalistic approach* to the understanding of organizational behavior. In this category are viewpoints that stress individual cognitive attributes, human experience, and individual differences as behavioral determinants. Some of these assume that employees act as individuals, while others assume that individual behavior is mediated by informal groups. In either case, however, the bias is in the direction of personalism or "psychologism" rather than "sociologism" or structuralism.

The fourth section deals with *integrative approaches* to organizations, which attempt to synthesize the various one-sided views into a single framework represented to be a more accurate model of organizational functioning.

Traditional Structural Approaches

Probably the most ambitious attempt to deal with the social-psychological aspects of the social structure was that of Marx (cf. 1887; or more particularly the collection of writings, 1964). Marx was concerned with properties of the social environment as they influence the characteristics of people who populated that environment. In this view, social stratification is seen collectively in the form of ideology; the stratification system is a way of organizing the social order.

Ideology, then, is an expression of the stratification system and develops after the stratification system is established.

Thus, for Marx, the individual is influenced entirely by the social order and has no important characteristics aside from the system of which he is a part. The only way to change people is to change the society.

Following Marx's lead, Durkheim (1902) maintained that once a society establishes a division of labor, the properties of that society can be defined independently of people. The characteristics of people can be examined as a function of the categories to which they belong. In other words, people can be characterized as a function of the structural properties of society.

Ironically, then, the earlier sociological writings of theorists such as Marx and Durkheim can be seen to provide a broad theoretical base for classical organization theorists who founded such schools of organizational thought as scientific management, administrative management, and bureaucracy. By implication the Marxes and Durkheims introduced questions of distinctions between organization theory and personality theory, whose solutions still tend to elude us (cf. Hunt, 1969b).

Scientific Management

In his scientific management, Taylor (1911) set out to study organizations with "scientific precepts." He was interested primarily in work performance, and his goal was to integrate properties of the man with properties of the job. To this end, Taylor worked out procedures, by means of time and method studies, that would utilize less energy and yield more work. This view (and its complements, e.g., Gilbreth, 1911) shows little concern with the worker as a person; rather, the worker was seen as an extension of the job. So far as the model was concerned, the worker simply had no life outside his job or separate from his tasks. The implications of this "ascribed status" have been illuminated by Daniel Bell (1956). And it is interesting to note in passing that many Tayloristic polemics are today reechoed in passionate debates about "computerism." What was written before about Taylor and scientific management is being written again about the inhumanity of the computer (cf., e.g., Simon, 1965).

In the Tayloristic model, work gets done by providing employees with work routines (or programs). There is no point, thought Taylor, in having men work at less than capacity because the work is poorly organized (the man-machine concept). The assembly line, therefore, can be run as fast as people can work, and time-method studies can be used to discover how fast people can work. In other words, experiments could be performed to discover the *one best method* or *best conditions* for doing the job.

Apart from a kind of Calvinist work-ethic and a trust in monetary inducements, Taylor had no motivational terms in his models. Motivation was largely assumed as an intrinsic moral matter. Furthermore, the "essential ability to work" was simply taken as given (cf. Neff, 1968, pp. 5–11, for a pointed discussion of this idea). But, if motivation and ability tended to be treated as intrinsic, organizational operation was viewed as dependent on extrinsic control.

For Taylor, the organization was an authority structure to be described in terms of *span of control*. This was the key to organizational design and created a system based on a centralized pyramidal authority structure. Taylor's plan was to devise a rational structure based on an authority hierarchy, with varying support functions (clerical, research, and development, etc.) where and when necessary—the line-staff notion. Thus, for Taylor, the guiding principle of organization was the span of control (or the authority structure), with components laid out in terms of vertical line and lateral staff.

According to Taylor, the organization could then be specified in terms of organizational charts (authority and line-staff) and job specifications. This was the organization, the *formal system*. Nothing else mattered. The organization was thus defined in terms of scientific principles used to describe the formal system.

Confined almost exclusively to the vantage point of top management, Taylor's view of the organization suffers from a severe lack of conceptual richness. Furthermore,

> Scientific management achieves conceptual closure of the organization by assuming that goals are known, tasks are repetitive, output from production processes somehow disappears, and resources in uniform qualities are available [Thompson, 1967, p. 5].

Administrative Management

The administrative management views fostered by Gulick and Urwick (1937) stressed the structural relationships among production, personnel, supply, and other service units of the organization. Here again the emphasis is on efficiency, and many Tayloristic principles are invoked. Efficiency is maximized by task specialization and the grouping of similar tasks into departments. Responsibility is fixed according to such principles as span of control or delegation and by insuring adequate control over employees so that action conforms to plans. The assumption in this view is that a master plan is known so that specialization, departmentalization, and control can be determined rationally. (Simon, 1957, for one, has criticized this view by enumerating the difficulties in attaining agreement with and constancy in master plans.)

Like its relative, scientific management, administrative management commonly ignores the organization's environment and boundary functions necessary to transact matters of organizational survival with that environment.

We might interject at this point the observation that reviews of scientific management concepts frequently convey a not altogether accurate sense of ancient history. To be sure, there have been sequences of appearance and development, but the scientific management movement is hardly dead, even if it appears more often now in the form of computer-oriented management science exercises in model building, operations research, cost-benefit analysis, etc. Moreover, it often disguises itself today in a more dynamic "systems" terminology and posture, sometimes leaving behind many of its earlier oversimplifications (cf. Haberstroh, 1965; Koontz, 1962). In anticipation, the same cautionary comment might be made about the human relations tradition (discussed below), which sometimes is painted as if it were solely a manifestation of the 1930s and 1940s. However, unlike Australopithicus (or Che), it lives (cf., e.g., Whyte, 1969—note especially p. 708).

Bureaucracy

Similar in many ways to the above views, bureaucracy (cf. Mouzelis, 1967; Weber, 1947) focuses on staffing and structure as a means of dealing with clients and disposing of cases—and the achievement of efficiency. In this view, efficiency is maximized by organizing offices according to jurisdiction and hierarchical position, appointing experts to offices, utilizing procedural devices for work situations, the promulgation of rules governing behavior of positional incumbents, providing differential rewards by office to provide incentive and patterns for career advancement (Quinn & Kahn, 1967; Thompson, 1967). As with other structural theories of organization, Weber felt that the proper use of rules, rewards, and punishments could render the human component negligible by divorcing the individual's private life from his role as an officeholder. Once more the significance of individual identity was uncomprehended: Indeed, the belief was widely implicit that "workers," unlike managers and other "good" people, had none.

The static nature of the organization as described by Weber inspired numerous eloquent attacks on bureaucratic theory. Notable among these were the criticisms suggested by Merton (1957) and Gouldner (1954), dealing with the unintended consequences of bureaucratic rule enforcement. Merton argued that reliance on depersonalized relations and strict enforcement of rules and jurisdictions in order to reduce human variability lead to rigid behavior on the part of position incumbents. This rigidity, based on the need to defend individual actions, leads to reliance on organizational precedent and use of the most easily invoked rule in dealings with clients. Often, however, the rule most easily invoked is least appropriate to the client's case. The client's challenge of the bureaucrat's judgment serves only to increase his drive to defend his actions and thence to increasingly rigid and inappropriate behavior.

In a somewhat similar vein, Gouldner (1954) pointed out an additional self-defeating consequence of reliance, stemming from demands for control, on the use of general and impersonal rules. He argued that the use of such rules offers little incentive for employees to do any more than the minimum amount of work that will avoid punishment. When management discovers that productivity has fallen behind goals, pressure is placed on supervisors to supervise more closely. This serves to increase the visibility of power relations within the work group and leads to even more dogged determination on the part of the workers to do as little as possible.

Thompson (1961) added that bureaucracies tend to attract monocratic types to supervisory positions who tend to enhance feelings of helplessness and insecurity in subordinates. Jasinski (1956) has found even more problems with the overreliance on efficiency controls. He specified seven: wasted time, higher maintenance costs, figure "fudging," low morale, impaired labor recruitment, higher unit costs, and lowered product quality. Schlesinger has provided an especially amusing account of bureaucratic blundering and job protectiveness in the United States Department of State in his book *A Thousand Days* (1965), and Bell (1956) has penned a scathing indictment of the Tayloristic-bureaucratic "cult of efficiency."

Other writers (e.g., Argyris, 1964; Bennis, 1969; Haire, 1963) have objected to bureaucratic theory because it is based on the assumption that man is lazy and not to be trusted. They further object to the accompanying moral notion of the intrinsic goodness of work and the punishment orientation implicit in classical theory. These critics agree that classical models of organization serve to stifle workers in their quest for "psychological success."

Argyris (1964), for example, has argued that to the degree that the dictates of classical theory are followed in an organization, the lower level employee will suffer because

> few of his abilities will be used . . . [he] will tend to experience a sense of dependence and submissiveness toward his superior because decisions of whether or not he works, how much he shall be paid, and so on, will be largely under his superior's control . . . [and] the worker will tend to experience a decreasing sense of self-responsibility and self-control [pp. 38–39].

Although most of the critics of classical organization theory view bureaucracy as either an ideal type or an undifferentiated lump, Hall (1963) has shown that the major discussants of bureaucracy cannot themselves completely agree on a list of definitive characteristics. Furthermore, in an empirical study, Hall demonstrated that organizations that are highly bureaucratized on one dimension may not be so on another. The dimensions studied by Hall were hierarchy of authority, division of labor, system of rules, system of procedures, impersonality, and emphasis on technical competence; other variables have been dealt with by Woodward (1965), Perrow (1967), and Harvey (1968) with comparable results. Clearly, then, bureaucracy exists in degrees, and there are very few so-called bureaucracies conforming in all ways to an ideal type. Nevertheless, most of the critical writing in this area tends to view it as an all-or-none phenomenon. The conclusion to be drawn is that classical organization theory is often used as a straw man or figurative foil for both critical virtuosos and theorists advancing alternative models.

Modern Structural Views

Sharing the same general rationale of organizational design as the traditional structural theorists—that the organization must be designed to fit the nature of people—the modern structural theorists, by altering their assumptions concerning people, have proposed alternative designs to bureaucracy. Instead of assuming as did classical theorists, that all workers are motivated by rational economic factors, modern structuralists undertake to view the worker as a total human being striving for self-improvement, self-expression, autonomy, recognition, or whatnot. It is further assumed that these strivings exist in equal amounts in all people but that their expression depends on the degree to which the social structure is able to accommodate such strivings.

The model used by this approach is nicely exemplified in Merton's (1957) reformulation of the concept of *anomie*. The problem for Merton was to explain why the frequency of deviant behavior varies in different social structures. He viewed the task as one of locating groups which typically engage in deviant behavior (or maladjustive behavior from the societal viewpoint) as *normal* responses to the social situation in which they find themselves.

Specifically, Merton argued that certain cultural axioms prescribing the means and ends of success have been well internalized by virtually all members of American society. He defined success as the attainment of wealth and stature, and the means to this end are hard, honest work and no horseplay. The culture, however, while vigorously stressing the ends, provides a social structure that severely limits access to the means. Anomie arises from the conflict between the goals that a culture sets up and the ineffectiveness of the prescribed means for attaining these goals. The incidence of this conflict is differentially observable in various strata of society, being more pronounced where an individual's position is weakened by the structure of society. Thus, people occupying positions in the lower social strata are more likely to be anomic as evidenced by criminal or ritualistic life styles employed as a means of coping with or defending against structural restrictions. The solution to problems of deviance then, of course, is to work at modifying the societal structure to allow for more freedom of legitimate goal attainment for incumbents of lower social strata (Gordon, 1964).

Several organization theorists have applied similar logic to social organizations. They maintain that all people have certain important common needs and that the structure of the organization interferes with the expression of these needs, especially for incumbents of positions low in the organization's hierarchy. This restrictiveness results in coping and adaptive responses such as minimal productivity, sabotage, low morale, and job satisfaction. Most of these theorists acknowledge the existence of individual differences, but they proceed to build their models as if these differences are of no account.

The immediate influences on these views rest with the Hawthorne studies (discussed later) and Maslow's (1954) theoretical formulations emphasizing man's inherent need to use his capacities and skills in a mature and productive way. Maslow specified a need hierarchy common to all normal people and consisting, in ascending order, of safety and survival needs, social needs, ego needs, autonomy needs, and self-actualization needs. A higher-level need cannot be satisfied until the needs below it in the hierarchy have been served. Maslow argued that it is universally true that all manner of human activity is an attempt to work upward in the hierarchy. Using this basic notion, modern structural theorists feel that since membership in a work organization is a central fact in men's lives, it is to the best interest of the organization and the people in it to alter the organization's structure to allow for the human quest for self-actualization.

Among the major advocates of this view is Argyris (1957, 1964), whose chief premise was that there is a lack of fit between requirements of bureaucratic organizations (the formal structure) and the needs of individual members to achieve "psychological success." Argyris wrote:

> our selection of the individual who aspires toward psychological success is not an arbitrary one . . . [T]here is an increasing number of psychologists who believe that self-esteem, self-acceptance, and psychological success are some of the most central factors that constitute individual mental health in our culture. If we are able to understand better how one may enhance the opportunity for individual psychological success, we believe that this will contribute toward individual mental health [1964, pp. 36–37].

As evidence for his premise, Argyris cited a number of empirical studies showing that workers at the lower end of the organizational hierarchy, as opposed to those at higher levels, suffer from poorer mental health, lower job satisfaction, lower levels of self-esteem, feelings of insecurity, and other related variables. The resulting feelings of dependence, subordination, and passivity on the part of the workers can only result in "frustration, failure, short-term perspective, and conflict [1964, p. 40]."

Further, Argyris took the position that an organization operated on classical premises suffers inefficiency to the extent that employees waste energy by engaging in activities of a compulsive, defensive nature. Thus, there is less human energy available for genuine productivity. As an outline for a solution, Argyris proposed a "mix model" in which "changes in the organization toward increasing opportunity for psychological success and self-esteem be made as long as we are able to show that they are decreasing the unproductive compulsive activities [1964, p. 147]." Argyris is quick to point out that this does not necessarily imply a people-centered organization because self-esteem is strongly coupled with responsibility to others. Furthermore, organizational changes designed to accommodate psychological success should continue only so long as the organization can more efficiently achieve its objectives, maintain its internal system, and adapt to its environment.

Argyris' view, then, contains a powerful belief in the reality of the forces exerted on individuals by an organizational structure. Since personality is considered a constant, individual differences can only be explained in terms of differential positional incumbency, and, indeed, the lower organizational levels are seen as the most restricting with respect to human need satisfaction.

McGregor (1960) embraced a similar view when he proposed "Theory Y" as a general solution to the problem of organizational improvement. Theory Y represents a set of assumptions about human motivation, based on Maslow's need hierarchy, which he felt should serve as a basis for organizational design. McGregor argued:

> Above all, the assumptions of Theory Y point up the fact that the limits of human collaboration in the organizational setting are not limits of human nature but of management's ingenuity in discovering how to realize the potential represented by its human resources. Theory X offers management an easy rationalization for ineffective organizational performance: It is due to the nature of human resources with which we must work. Theory Y, on the other hand, places problems squarely in the lap of management. If employees are lazy, indifferent, unwilling to take responsibility, intransigent, uncreative, uncooperative, Theory Y implies that the causes lie in management's methods of organization and control [p. 48].

In other words, the structure of the organization is responsible for difficult employees, and an alteration in this structure is called for.

Other theorists have, in whole or in part, fostered a similar point of view. Likert (1961, 1967) did so by arguing that "System 4" (trust and participation) is the only leadership style which can yield maximum effectiveness. Gellerman (1963) posited competence and power as general motives and suggested that the structure take these into account. And Blake and his co-workers (cf. Blake & Mouton, 1968) also proposed a "one best" route to managerial effectiveness—the 9–9 theory.

The solutions proposed by the several writers holding to modern structural viewpoints suggest reordering organization away from bureaucratic designs so that they will allow for participative decision making, face-to-face work groups, mutual confidence between superior and subordinate, job enlargement, increased responsibility at lower positions, and so on.

The modern structural theorists have met with criticism on two main grounds: Their global personality assumptions and their propensity to blame the structure of the organization for all of its ills. Examples of the first type of criticism have emanated from the pen of Strauss (1963, 1968). Strauss felt that the notions of such people as Argyris (1959), Herzberg, Mausner, and Snyderman (1959), Maier (1955), Maslow (1954), and McGregor (1960) are laden with an inner-directed academic bias in which righteous professors impose their own values on members of population segments who may have quite a different idea of self-development. J. K. Galbraith (1967), however, felt that modern structural theorists seem to feel that if workers gain most of their satisfactions away from their jobs, then the organization has somehow failed in both efficiency and elementary humanism and must, of necessity, be restructured. Strauss (1963) continued by asserting that the injudicious use of power-equalization practices may actually adversely affect the security needs of those workers who are not prepared for the added responsibilities that accompany an expanded role. Indeed, a closely related phenomenon figured prominently as a source of disturbance in Levinson, Price, Munden, Mandel, and Solley's Midland Utilities study (1962). Further, power-equalization techniques may be more expensive to employ than they are worth in highly programmed situations where simply adequate performance or keeping up with the assembly line is all that is required (cf. Perrow, 1967, for a discussion of this and related matters in a wider context of reviewing relations between technology and organization). Moreover, like their traditional counterparts, the modern structuralists are susceptible to criticism for their advocacy of "one best way" notions, albeit this time applied mostly to managerial methods (cf. Fiedler, 1965).

Sayles and Strauss (1966) maintained that the personality versus organization model is most applicable to employees who do highly skilled or creative work in large companies (e.g., research and development scientists, industrial designers—people similar to the model builders). For the workers who must perform jobs that are more difficult to make intrinsically interesting, Sayles and Strauss (1966) predicted a reemergence of the kind of leisure-based culture characteristic of ancient Greece and Rome. It is too late, they feel, for our society to revert to a production system based on individual craftsmanship. Katz and Kahn (1966) have pointed out, for instance, that a good many mass-produced automobiles can be bought for the price of one Rolls Royce. As noted above, Galbraith (1967), too, has voiced similar sentiments, although Bell (1956), while acknowledging the trend, has expressed skepticism as to whether there can be adequate substitutes for loss of satisfaction from work.

At any rate, a challenge to the structural suppositions of the human relations theorists has come from an extensive review of the empirical literature by Porter and Lawler (1965). These authors studied the literature pertaining to seven aspects of organizational structure: Organizational levels, line and staff hierarchies, span of control, subunit size, total organization size, hierarchical steepness, and degree of centralization or decentralization; they concluded:

> Five of the seven properties of organization structure (span of control and centralization/decentralization being the two possible exceptions) have been shown to have some kind of significant relationship to either job attitudes or job behavior, or to both of these types of variables. However, . . . experimental "proof" of cause–effect relationships between structure and employee attitudes and behavior is elusive and almost nonexistent [p. 47].
>
> [T]here are already enough indications in the literature to support a greater research effort to investigate the interactions among structural properties of organizations in their relationship to employees' job behavior and attitudes. Too much previous theorizing in the area of organizations has neglected such interaction possibilities and hence, there has been an unfortunate tendency to oversimplify vastly the effects of particular variables. Organizations appear to be much too complex for a given variable to have a consistent unidirectional effect across a wide variety of types of conditions [p. 48].

Similar sentiments relative to the contingent nature of organizational processes have become commonplace in the literature. And their pertinence is underscored by recent work on determinants of formal organization structure and the impact of formal structures on member characteristics (cf., e.g., Hickson, Pugh, & Pheysey, 1969; Whyte, 1969, especially Chapter 3). Structural properties of organizations surely "seep into" their members, as Udy (1965) stated, but the reverse is also true, and, in any case, the relations usually are not simple.

Other criticisms have also been leveled against the human relations structuralists. Pugh (1966) has noted that the theorists reacting against the classical school seem to have an implicit bias against formal organization. Leavitt (1962) felt that much organizational restructuring based on human relations philosophies has been done on the principle of defeating Taylorism and has too often thrown out the baby with the bath water. Leavitt felt, in particular, that the overreaction against classical theories has served to impede the salutary introduction of principles from communication and information theory into organizational design. Hunt and Lichtman[2] have cautioned managers against giving their employees the impression that productivity is somehow not important; in the first place, it is not true; second, it only serves to confuse employees and may create stress in those places were a human relations orientation is supposed to relieve it. And Katzell (1962) has cited evidence of conventional organizations that seemed to be functioning perfectly well and of human relations programs that were rejected by workers.

---

[2] Hunt, R. G., & Lichtman, C. M. Concepts and strategies for training work supervisors as employee counselors. Buffalo, N. Y.: State University of New York, 1969. (Ditto)

It would seem, then, that current structural theories with emphases on human relations have made many of the same errors of theoretical presumption as did the classical theorists against whom they were ostensibly reacting. Notably, their assumptions of global personality characteristics give their theories the same one-sided, closed system, quality of the classical theories. They have a way of proposing the one *best* solution to the efficiency of all organizations and the *one best way* to motivate all members. Their view then is not only simplistically one-sided, but unidirectional: Effects flow almost entirely from the organization's properties to the individual. As compared with modern varieties, traditional structuralism might, however, be said to hold a somewhat friendlier attitude toward organizations and a somewhat less humanistic one toward its members. Traditional structuralists tend to see organizations as societal benefactors needful only of augmented efficiency, whereas the modernists look upon the organization as a more oppressive agency suitable for reform. In either case, however, the belief prevails that apt structuring of the features of the organization will allow the individual's natural inclinations to issue forth as more effective performance. (And Taylor was as convinced as anybody that this would be accompanied by happier workers.)

In his discussion of "work as a psychological problem," Neff (1968), however, argued that it is evident that people exist who cannot (or will not) work effectively under even the best of conditions. The question he raised under the rubric of the "work personality" is, in essence, What psychological characteristics are requisite to entry and performance in "work roles"? (cf. also Dawis, Lofquist, & Weiss, 1968).

Personalistic Views

Although global theories of personality have been of benefit to organization and management theory—especially their implications for organizational design—they are often difficult to study in a precise empirical manner, and they do not provide sufficient differential predictive power to be useful by themselves for an understanding of organizational behavior. As a result, many students of organizations have chosen to concentrate their theoretical and empirical efforts on the discovery of the influence of more particularized measures of personality. Schein (1965) commented:

> Man is not only complex, but also highly variable; he has many motives which are arranged in some sort of hierarchy of importance to him, but this hierarchy is subject to change from time to time and situation to situation; furthermore, motives interact and combine complex motive patterns (for example, since money can facilitate self-actualization, for some people economic strivings are equivalent to self-actualization). Man is capable of learning new motives through his organizational experiences . . . and . . . is the result of a complex interaction between initial needs and organizational experiences. Man's motives in different organizations or different subparts of the same organization may be different. . . . Man can become productively involved with organizations on the basis of many different kinds of motives . . . [p. 60].

Many of the same points have been made by Hunt (1969a) in discussion of the role of profit in business affairs. In any event, the general assumption of the personalistic view is that people react to their organization on the basis of their perceptions of it. These perceptions are based on people's needs, motives, and values. Therefore, to understand human behavior in organizations, one must understand how individuals differ with respect to personalistic variables; to change organizations, one must alter the perceptions of people. The impetus to this view rests largely with the writings of Kurt Lewin and the Hawthorne group.

Lewin's field theory and his group dynamics work (cf. Lewin, 1951) prefigured more than one important contemporary pattern of organizational analysis. The group dynamics research, of course, was one of the roots of T-grouping, while field theory per se constituted one of the early gestures in psychology toward a systemlike interpretation of individual and group behavior. Among other things, for instance, Lewin argued that action was no simple outcome from mechanical stimulus-response linkages, but was a complex resultant of coacting mutually influential elements coalescing as dynamically shifting momentary fields of force in a psychological environment, the life space.

Broadly phenomenological in orientation, Lewin held the characteristics of the life space to be indefinable apart from the cognitions of the person-actor; overt behavior was construed as essentially derivative from fields of force consequent upon modes of organization of that cognitive space. Through his cognitive contributions, the actor is thus viewed as active in the behavioral process and not as just a passive recipient of discrete stimuli. Indeed, in the Lewinian scheme, although constrained ecologically, the frame of behavior was phenomenal; the life space was a perceptual or experiential creation. Furthermore, little room was left for examining the accuracy of perception. The field of behavior was, so to say, principally "intrapsychic."

Now, if an organization has no properties aside from people's perceptions of it based on their needs, values, and attitudes, it can have no real structure because it cannot exist independently of its particular time-relative membership (cf. Hunt, 1968). The process of organizational analysis must then consist largely of personnel assessments and identification of ad hoc liaisons; organizational change must consist solely in the changing of people. Such a premise is fundamental (if implicit) to what may be termed "clinical" approaches to organization development (e.g., those associated with T-group perspectives; cf. Schein & Bennis, 1965).

About the same time as Lewin was writing, the Hawthorne or Western Electric studies (Mayo, 1933; Roethlisberger & Dickson, 1939) were being conducted. Although originally conceived as a test of Taylorism—a search for the one best level of illumination, among other things—the findings indicated that there was no one best level. Rather, workers stepped up production at all levels. The Hawthorne studies became important principally because the investigators went on to ask *why* there was no relation between illumination and output, and the studies conducted to discover the answer led to neoclassical organization theory.

From their studies the investigators concluded that friendship patterns were the heart of organization. They subsequently advocated human relations as a set of techniques by which people could be motivated. By maintaining that the organization is what is perceived by the employees to be the case, the proponents of neoclassical organization theory abandoned the formal structural notions of the classical theorists. The important system was no longer the formal organizational structure, but the informal relations at all organizational levels. What had traditionally been called the formal structure, then, was nothing but the manifestation of the workings of the informal system (or else somebody's idealization of it).

As the Hawthorne researchers saw it, however, the implications for management did not lie in any gross restructuring of the organization, but rather in a program of individual counseling designed to change the perceptions of individual workers toward the organization. The classic studies of Coch and French (1948) and Lewin (1958) were direct offshoots of this line of thinking. In both of these studies the persuasive potential of informal group membership was used as a vehicle for changing people's perceptions of the structure *as it is;* there was no intention of changing the basic structure of the system, there was only the attempt to get people to accept it. Although this new approach came to be called participative management, the fact was that management had no intention of altering the organization—only of changing people's attitudes toward work. Among others, Bell (1956) has been vocal about the "manipulative" qualities of human relations operations, and recently even Bennis (1970) has commented on this tendency.

Individuals and Informal Groups

The Hawthorne researchers have been subject to sweeping critique. Carey (1967) has argued that their conclusions neither followed from nor were they supportable from their data; if anything, they demonstrated exactly what they maintained had been disproven. Yet, be that as it may, few more influential works could be found—they ushered in an era. But as Katz (1965) has pointed out, the Hawthorne studies led to the demise of the notion of economic man without, however, actually making social organization the focus of analysis. The Hawthorne studies argued that man is motivated by a good deal more than monetary self-interest; indeed, that the worker's peer group could actually guide a man away from his own self-interest by establishing a norm of restricted output. Although the group factor in work situations was emphasized, and the concept of social system introduced, an individual focus was retained. The function of the group, in this view, was to relieve the boredom of the individual.

Roy (1960), however, has argued that the function of the group is a good deal more than that. The work-group culture reaches outside the organization by defining the social reality not only of the work situation but also of the worker's place in the larger society. The work group as an informal structure, argued Roy, provides a continuity between life outside and inside the organization. This general position has received wide support in the literature (e.g., Bennis & Shepard, 1965; Festinger, 1950; Walker, 1958).

Likert (1961, 1967) typifies the group dynamics approach in his advocacy of group decision processes at all levels. He proposed a type of organizational structure consisting of many overlapping groups that extend across adjacent management levels. Each foreman, for example, would be a member not only of the work group but also a full-fledged member of the next higher supervisory level. Thus, the supervisor in one group is a subordinate in the next, and so on, at successive levels. Horizontal as well as vertical linkages are built into this system. The emphasis is clearly on groups rather than on individuals. "An organization will function best when its personnel function not as individuals but as members of highly effective work groups with high performance goals [Likert, 1961, p. 105]." Likert felt that employee participation in decision making, when accomplished according to his "linking pin" notion, will not reduce supervisory influence (a view shared by Tannenbaum, 1966), but rather will serve to change the attitudes of work-group members to better conform to organizational goals. Moreover, like others of its genre, Likert's solutions to organizational problems rarely carry important implications of a basic structural order; indeed, by and large they presume a generally bureaucratic system and speak mainly to operations within it (cf. Bennis, 1970).

Despite contemporary recognition of the influence of *informal groups* on the formal organization, many modern theorists have preferred to limit the domain of their analysis to individual experience. Vroom (1964), for instance, has recently applied Lewin's general theory of personality to the study of organizational behavior, hypothesizing performance to be a function of the worker's perception of the abilities required by the job, the degree to which the person perceives himself as having these abilities, and the degree to which he values the possession of such abilities. Vroom, thus, stressed the affective consequences of the degree of consistency beween a person's performance and his self-concept. In this view a person is motivated to perform effectively when effective performance is consisent with his conception of his abilities and with the value he places on them. Porter and Lawler (1968) have similarly hypothesized that managerial performance is a function of the perceived value of the reward, the probability that effort will bring the reward, and the accuracy of role perceptions. Smith and Cranny (1968) have presented a similar formulation, and March and Simon (1958) and Cyert and March (1963) have also stressed personalistic variables in their individual decision-making models of organizations. They have used such expressions as "the perceived consequence of evoked alternatives," "the expected value of reward," placing the locus of the organization within the individuals who populate it.

Katz (1964) has posited that the motivation to produce and motivation to belong are orthogonal dimensions, and this indepenence helps to explain the corresponding lack of relationship between job satisfaction and productivity. Bass (1965) argued that it is not necessary to study workers with reference to any basic needs or tensions; rather, there is just as much predictive power to be had by restricting research to "organizationally relevant" motivational dispositions. Task orientation, self-orientation, and interpersonal orientation are the three proposed by Bass. These and some related ideas have been well-summarized by Quinn and Kahn (1967).

In his book *Eupsychian Management*, Maslow (1965) took a more radical view by arguing that work in itself has a kind of symbiotic quality. Satisfaction can indeed be gotten out of good work, but this is not a social or socially derived (ego) need. Rather, work is a primitive need independent of social need satisfaction and can be intrinsically interesting. In principle, Maslow's view is not altogether unlike that of Bell (1956), which we have cited, or of Herzberg et al. (1959), who have claimed that the only true job satisfaction can come out of work itself. Good working conditions and satisfactory social relationships around the job can only serve to make the job tolerable.

A review of the literature concerned with studies of individual worker differences was written by Smith and Cranny (1968) and provides further evidence for the value of such research. In sum, recognizing the inadequacies of classical organization theory, many writers have turned the coin and concentrated their theoretical efforts toward evaluating the role of the individual and the work group as a determinant. They have also encouraged an individual differences approach to the study of organizational behavior, by proposing intrapersonal dimensions along which people may vary. An especially clear illustration of this can be found in Hackman's recent (1969) reworking of the Herzberg two-factor notions about work motivation.

But this view is no less one-sided than the structural approaches because it has little to say about the formal organization. Any comprehensive theory of organizational behavior, of necessity, has to integrate the formal structure, the processes occurring within individuals, and the function of informal groups (cf. Hunt, 1968; Perrow, 1967).

The solution, according to Scott (1961), lies in the remembrance that the main problem with the structural and personalistic views was not one of inacurracy but rather incompleteness and shortsightedness. In neither approach is there adequate integration among the numerous facets of human behavior studied by it. The task of the modern organization theorist, therefore, is to combine the several views into a more accurate model of organizational functioning.

Mechanic (1963) has championed the use of the multivariable or "eclectic" approach to organizational studies, as opposed to "one-sided" approaches committed to a single view of human nature. He maintained that one-sided approaches tend to encourage strong partisanship to a particular view and thus engender misunderstandings, communication difficulties, unjustified attacks, while they also discourage interdisciplinary cooperation.

> However, if "one-sided" analysis is defined openly as a strategy of approach rather than a dogmatic assertion of priority, the dialogue between theories may be improved and communication and cross-fertilization may be preserved at the same time . . . [W]hile it may be valuable to urge antieclecticism in theory building, it is essential to urge eclecticism in the use of research approaches and methodologies. Perhaps the most significant recent trend in organization studies is the utilization of various modes of data collection in the same study [Mechanic, 1963, pp. 172–173].

The virtues of the one-sided versus the eclectic approach depend, in part, on the purposes of the theorist. If his purpose is to convince bureaucratic managers of the advantages of human relations practices, there is probably no harm in taking a more extreme personalistic view of human behavior in organizations; presumably, the target is already familiar with the operations of formal rules. If, however, the theorist's intent is to build an accurate and workable general model of organizational functioning, then it is clear that none of the one-sided views proposed to date can do that job. The next section deals with some approaches to the discovery of a general solution to the organizational dilemma.

Integrating Approaches

Increasing disenchantment with the myriad of one-sided and normative approaches to organization theory has led many writers to favor the use of alternative models. Such models typically represent attempts to integrate existing approaches and/or reflect the influence of situational factors on organizational functioning. Katzell, for example, wrote:

> by concerning themselves with certain preselected dependent variables . . . and not with others that are relevant to the system, by relying heavily on unproven assumptions, and by their tendency to generalize rather than paying due attention to surrounding circumstances, such theories typically take the form of a blueprint or master plan which is prescribed as the one best way to organize and manage all work [1962, p. 104].

Katzell felt that many factors must be taken into account before it can be decided what organizational policies and practices are likely to work best. As a starting point, he suggested five such parameters:

1. The first parameter mentioned by Katzell is size, or the number of interdependent organizational members. Hickson et al. (1969) have also given prominence to size factors in their analyses of determinants of structure. Indik (1963) reviewed the literature relating organizational size to member participation in terms of absence and turnover (a finding which calls to mind Barker's ecological studies of schools, cf., e.g., 1963). In another paper, Indik (1965) interpreted these findings by positing that larger organizations contain more potential and necessary communication linkages among members, rendering adequate communication more difficult to achieve. The result of inadequate communication among members serves to decrease interpersonal attractiveness and, in turn, member participation rates. Porter and Lawler (1965) also reviewed the literature on organizational size and concluded that although job satisfaction and morale tend to be lower in large organizations, the general findings relevant to this area are neither reliable nor clear. These authors felt that perhaps subunit size is a more crucial determinant of employee reactions than is total organization size. Woodward (1965), too, has pressed for a more differentiated concept of size, stressing the importance of indexing size by reference to the size of managerial subsystems. There is clearly a need for more research in this area.[3]

2. The degree of interaction and interdependence is another parameter suggested as important by Katzell. Katz and Kahn (1966) have theorized that the degree of task interdependence is an important consideration in choosing the appropriate authority structure for an organization. They argued that where such interdependence is high, where creative requirements are minimal, and where identification with organizational goals is not required, a hierarchical authority structure, with its implication of close supervision, is indicated. Where the opposite circumstances obtain, a more democratic alternative should be used for maximum efficiency. Vroom and Mann (1960) have proposed that differences in work-group task interdependence account for the choice of autocratic supervisors by some groups and democratic supervisors by others.

More generally, Hackman[4] has shown the importance of task structure to patterns of group interaction; Sommer (1967, 1968) has demonstrated the pervasive influence of broader ecological factors in group (and, by inference, organization) processes, and W. F. Whyte (1969) may be consulted for a wide-ranging discussion of various aspects and implications of "the structuring of the work environment" (cf. especially Chapter 3) Hunt (1970) has discussed at length the organizational significance of task uncertainty and has identified distinctive modes of structural adaptation to it. Relatedly, Haberstroh (1965) provided a review of the role of information systems in task modeling and organizational design. He noted, for instance, that "the coherence and unity of an organization is intimately related to its information system [p. 1192]." Therefore, he maintained, variations in such systems will precipitate "major [organizational] repercussions." Haberstroh promoted the idea of close and explicit integration of information systems and task models (cf. also Cleland & King, 1968; Glans, Grad, Holstein, Meyers, & Schmidt, 1968).

---

[3] Recently Blau (1970) has published a systematic theory relating organization size and structural differentiation that could prove a productive guide for such research. He touches on various nuances of both dimensions, but his "basic generalizations are (*a*) increasing organizational size generates differentiation along various lines at decelerating rates; and (*b*) differentiation enlarges the administrative component in organizations to effect coordination [Abstract, p. 201]."

[4] Hackman, J. R. Functions of group interaction. Paper presented at Colloquium of the State University of New York, Buffalo, October 1969.

3. Another variable to be considered is the personalities of organization members. Their motivations, expectations, abilities, and other personal qualities seem to be of paramount importance in determining policies and practices of organizations. Employee satisfactions and attitudes have been linked to subcultural differences (Katzell, Barrett, & Parker, 1961) and cross-cultural differences (French, Israel, & Aas, 1960), as well as to individual difference measures (Smith & Cranny, 1968) and patterns of immediate supervision (Ronan, 1970).

4. The degree of congruence or disparity between organizational goals and the needs and goals of the members is a fourth factor to be considered as a determinant of organizational design. Katz and Kahn (1966) distinguished between two types of goal internalization that might be aroused in an employee. The first of these occurs when the employee identifies with the job itself and is, therefore, conducive to high quantity and high quality role performance. Job identification is exemplified by the craftsman or scientist who delights in the work he does and is rewarded by the opportunity to express his abilities and to make his own decisions. Although morale and satisfaction are well-linked to the variety and challenge of the job and the gratifications that accrue to workers, such individual craftsmanship is infeasible and expensive to apply to many positions in a company concerned with the mass production of consumer goods and also in some service operations. One solution is to eliminate dull assembly-line jobs through automation (cf. Bell, 1956, for a useful general discussion of this issue). Mann and Hoffman (1960) and Blauner (1964) have argued that after automation has occurred, the remaining jobs provide more freedom, responsibility, and dignity. The result is a decrease rather than an increase in worker alienation.

The second type of identification proposed by Katz and Kahn occurs through internalization of organizational goals rather than through the job itself. This type of motivation is associated with value expression and self-identification. In industrial organizations this type of internalization is usually found among management personnel, although in voluntary organizations it may extend to the rank and file. Katz and Kahn argued, however, that complete internalization of organizational goals is less common than two types of partial internalization. The first of these concerns general organizational purposes that are not unique to a particular organization. A scientist, for example, may internalize the research values of his profession and, to the extent that the organization that employs him shares similar goals, coincidence in worker and organizational goals will occur. However, the scientist could work in many other organizations just as easily. The same, of course, could be said of teachers and other professionals whose attachments to particular organizations are derivative and chiefly as a simple locus of occupational activity.

The other subtype proposed by Katz and Kahn is the internalization of subsystem goals and values. Since one's own organizational unit is often more visible to him and is frequently his immediate source of reward, it is often easier to internalize the smaller unit goal than that of the entire organization. This often leads to the aggrandizement effect (Caplow & McGee, 1958; Morane, 1967), that is, tendencies to rate one's subsystem higher than outsiders do. This may lead in turn to empire building, parochialism, petty jealousies, and factionalism within the organization. Thus, depending on which tendencies are involved, a high degree of goal internalization on the part of employees may not always be a bed of roses for the organization; but, in general, it is clear that the degree and variety of employee identification can have a profound influence on organizational policy and practice.

5. Katzell's final parameter has to do with who in the organization has the necessary

ability and motivation to take action that will further its objectives. Rubin and Hunt (1969) have discussed this matter with special reference to what they called "mission identification." After reviewing some of the literature on supervision, Katzell observed that

> the best way to organize and manage work is heavily conditioned by those in the organization who have the knowledge and motivation to get the job done. More specifically, it would seem necessary to mold the organizational system so that it maximizes freedom of action and initiative for those who can and will take effective action, while eliciting compliance, support, or noninterference from others. Depending on the organizational loci of expertise and dedication, the appropriate work system may vary widely from situation to situation [1962, p. 107].

Katzell concluded that "the sheer number of variables that can be used to describe organizations is staggering [1962, p. 105]" and that ways must be found to work with them meaningfully (and we have hardly even mentioned technological variables; cf. Hunt, 1970). One way to do this, he proposed (not surprisingly), is to discover the genotypic dimensions that underlie phenotypic complexities.

It is the recognition of the vast number and complexity of organizational variables that has been responsible for a decided shift in emphasis on the part of many organization theorists over the past decade or so. Perhaps it is more accurate to say that recognition of the complexity of the facts of social structure and social organization was responsible for the belated development of the psychological approach to the study of organizations. Until perhaps two decades ago the social sciences lacked the conceptual tools for dealing with many important facets of social life (Katz & Kahn, 1966). With the introduction of the biological notion of an open system (von Bertalanffy, 1950) into the social sciences (Miller, 1968; Morane, 1967) came a practicable possibility for integrating the plethora of one-sided views of social organizations.

Social Systems

Among the earliest and most generally useful systemic theories of organizational behavior was the theory put forth by Homans (1950). Homans commenced by positing that any social system exists in a three-part environment: a *physical* environment (geography, climate, etc.), a *cultural* environment (the norms and values and goals of the society at large), and a *technological* environment (the state of knowledge and instrumentation available to the system). At the next level the social system itself has certain requirements and goals that are translated into specified activities and interactions for members of the system. The behavior required by the system or determined by its environment is called the *external* system by Homans. For example, in a work organization, management constitutes a large element of the external system by making decisions that bring certain people together in particular ways, for instance, through job specifications, work methods, prescribed layouts, and selection of personnel.

Although the people brought together by the management may be strangers at first, they eventually come to know one another, develop cliques and friendships, socialize on and off the job, help one another in their work, agree to restrict production, and so on. They now have another basis on which to associate, one that modifies and influences their behavior over and above (and perhaps instead of) the effects of the external system. The new form of behavior comes about as a reaction to the demands of the external system. Homans called this phenomenon the *internal system*.

Homans further specified that the elements of a social system can be sorted into three categories which contain aspects of both the internal and external systems: *Activities* (the things people do, the acts they perform), *interactions* (activities that link people together so that the activity of one person has an effect on the activity of another), and *sentiments* (internal psychological states, e.g., emotions, feelings, beliefs, values). Homans postulated that activities, interactions, and sentiments are mutually dependent on one another so that a change in any one will produce a change in the others. For example, positive sentiments between two people lead to increased interaction, or vice versa. Homans further argued that the internal and external systems are mutually dependent; for instance, technology influences interaction, which affects informal relationships. Conversely, the Hawthorne studies demonstrated how the informal system influenced "formal" production quotas.

Lastly, the two systems and the environment are mutually dependent. The environment may change the formal organizations (e.g., through federal and state legislation, union activities), which, in turn, will serve to alter informal organizational relations. Contained in Homans' formulations is the explicit recognition that a social organization, at any point in time, is the outcome of a pattern of interactions between the organization's stated requirements, its environment, and the characteristics of the people who populate it. It ties the emergent system of people's actual everyday behavior to organizations and the external system of formal plans, culture, and other groups that mold emergent behavior.

A more recent comprehensive discussion of organizations as open social systems is *The Social Psychology of Organizations* in which Katz and Kahn (1966) give precision to many concepts of organization by including them as part of a formal model within which a concept is defined partly in terms of its relations to other concepts.

At the outset of their book, Katz and Kahn discuss nine common characteristics of open systems:

1. Open systems, like biological organisms, must import some form of energy from the environment in order to survive.

2. Open systems contain a through-put process by which the imported energy is transformed; the through-put of an organization may be the creation of a new product, the providing of a service, the modification or treatment of human beings, etc.

3. "Open systems export some product into the environment, whether it be the invention of an inquiring mind or a bridge constructed by an engineering firm [p. 20]."

4. An open system consists of cycles of events. The product exported to the environment furnishes sources of energy renewal for the input so that the cycle may be repeated. It is important to note in this context that the structure of a social system

> is to be found in an interrelated set of events which return upon themselves to complete and renew a cycle of activities. It is events rather than things which are structured, so that social structure is a dynamic rather than a static concept [p. 21].

5. Open systems are further characterized by negative entropy. An open system imports more energy from its environment than it expends; it can thus store energy to counter entropic forces.

6. Energic inputs, which become transformed or altered by the through-put, are not the only form of system inputs. Information inputs, negative feedback, and the coding process are inputs of an informative character and serve to provide the system with information concerning its own functioning in relation to its environment.

7. An open system is characterized by a steady state and dynamic homeostasis. This does not imply a motionless or true equilibrium, but rather a force that seeks to preserve a constant ratio between the parts of the system. Thus, when one element in the system changes, forces are exerted to preserve the character of the system by the proportional alteration of all other elements. Thus, as with living organisms, organizational evolution is a symmetrical process.

8. Open systems become more differentiated and elaborated over time. Diffuse global patterns give way to more specialized functions.

9. Finally, open systems are characterized by the principle of equifinality. There is no *one best way* for a system to reach a given final state from a particular initial state. Contrariwise, given similar initial states, open systems may reach quite different final states. In short, there are a variety of paths between any two given points in a system's existence.

The underlying notion of open-system theory is that everything that does or can happen is dependent on everything else that does or can happen; that is, all events are correlated. Thus, the concept of cause and effect is discarded. While a closed system uses up all its energy in the through-put process and becomes simpler over time as a result of entropic processes, an open system has a permeable boundary that allows it to draw sustenance from its environment. Newcomb (1959) has written:

> I have chosen to emphasize "system properties" rather than single variables which contribute to them, and consequently none of the variables has an enduring status either as independent or as dependent . . . a change in one system variable is likely (under certain conditions) to be followed by a specified change in another system variable, but according to others a change in the second is a precondition for a change in the first [p. 388].

All elements of the system change when the value of any one is changed. That is, there is a redistribution of energy through some form of equilibration process. An open system is, therefore, dynamic; it is expending energy, and, as a result, it is in process and ongoing. (Doubtless the most wide-ranging treatment of system concepts and applications will be found in Buckley's 1968 compendium.)

The open-system model has also been vigorously advanced by the Tavistock group (e.g., Rice, 1963; Trist, Higgin, Murray, & Pollack, 1963) who have combined it with the less general concept of the sociotechnical system. The purpose of the sociotechnical concept is to bring into the same framework both the ideal efficiency of the system and the practicability of attaining the ideal with human beings and realistic conditions of work. Any productive organization, in this view, is the result of the interaction between technology and the system of relationships between those who perform the jobs (the social system). Each system influences the other (cf. Hunt, 1970, for a detailed discussion of these relations). This notion, of course, is not unilike Homans'.

However, the Tavistock group, by adding the notion of a sociotechnical system to the open-system model, served thereby to add recognition of the importance of the channels of interaction between the organization and its environment. Thus, the organization suffers constraints from two sources. On the one hand, the environment controls raw material, labor, money, and so on, all at the input side, while the output of the organization depends on such variables as consumer preferences, the state of the economy, and government regulation. And if that is not enough, additional constraints are imposed on the organization by the expectations, values, and norms of its human incumbents.

Of utmost importance in this model, however, is the notion that human capacities, preferences, and expectations are not necessarily personalistic elements, but may be influenced by experiences within the organization. Thus, the functions of personnel selection and organizational design are of *equal* importance—human behavior in organizations is the result of the interaction between formal role requirements and the nature of people (cf. Hunt, 1967).

In a more recent volume from the Tavistock Institute, Miller and Rice (1967) have employed a boundary model as a framework for examining problems of organizational control (see also Tannenbaum, 1968). The focus of this model is to make a distinction between the boundaries of the activity system ("that complex of activities which is required to complete the process of transforming an intake into an output [p. 6],") task group (comprising those individuals employed in an activity system), and sentient group ("being the group to which individuals are prepared to commit themselves and on which they depend for emotional support [p. 253]"). Miller and Rice presented various organizational models that represent different degrees of coincidence between task, sentient, and administrative boundaries and evaluate each in terms of their ability to control task performance, to ensure the commitment of members to the organizational objectives, and to regulate relations between task and sentient systems.

One of the major hypotheses posited by Miller and Rice is that in organizations in which task and sentient groups coincide on a more or less permanent basis, only short-term effectiveness can be achieved. "In the longer term such groups can inhibit change and hence can lead eventually to deterioration of performance and, in consequence, to social and psychological deprivation rather than to satisfaction [p. 253]." As an alternative, Miller and Rice promoted temporary project systems as a general solution. These authors felt that such temporary work groups (such as an airline flight crew) will discourage the formation of group standards detrimental to organization efficiency, while providing sufficient social relationships to satisfy relevant employee needs. Bennis and Slater (1968), of course, have made of such ideas as these a basis for general societal design.

In a related vein, Crozier (1964) has argued that for a given organization there is no such thing as *a* structure; rather, each organization may best be characterized as an aggregate of structures, for instance, a power structure, a communication structure, or a friendship structure. Thus, an organizational member may be seen as occupying a variety of positions with respect to the various structures in which he is included. In Crozier's view, such dimensions as work flow and dependence, as well as those mentioned above, may reveal numerous relationships not accounted for by the formal structure.

The recognition by the open-system model of the legion of variables and relationships affecting the behavior of social systems need not, however, imply that such phenomena are randomly distributed throughout an organization; they tend to be specialized by location. For example, Parsons (1960) has proposed that organizations exhibit three distinct levels of responsibility and control—technical, managerial, and institutional.

In his view, an organization contains substructures that exert forces to maintain a specific category of functions necessary to organizational survival. The technical or production structures are primarily concerned with the through-put of the organization—the processing or conversion of the input—and strive to achieve technical proficiency. The managerial structure services the technical subsystem by procuring necessary resources for the input and by mediating between the technical structure and the environmental elements to which the product will be exported. The managerial level controls or administers the technical substructure (not necessarily unilaterally or even unidirectionally, it should be noted) by deciding such matters as what the business of the organization shall be, the scale of its operations, and employment and purchasing policy.

Finally, the organization itself is part of a wider social system, which is the source of higher level support for the organization's goals. Although an organization may be relatively independent with regard to its formal controls, it must depend on its environment for meaning, legitimation, and resources. The overall responsibility for maintaining a favorable environment for the organization lies at the institutional level.

Parsons further argued that at each of the two points of articulation between the three levels there is a qualitative break in the simple continuity of line authority because the functions at each level are qualitatively different. The difference is one of kind, not degree. Organizational survival consists of the adequate functioning of these three levels as well as adequate interaction. If any level withholds its contribution, the system is dissipate.

Katz and Kahn (1966) have expanded Parsons' notion by offering a somewhat finer articulation of basic organizational substructures. Production structures are concerned with the work that gets done; supportive structures procure the input, dispose of the output, and deal with institutional relations; maintenance structures promote adequate role performances; adaptive subsystems serve to change the organization in line with a changing environment; and managerial subsystems direct, adjudicate, and control the several subsystems and activities of the organization.

The upshot of this line of thinking is that since these basic substructures are interdependent, a change in any one will bear implications of change for the others: For instance, differences in technology will make for differences at the managerial, supportive and other levels; differences in environment will, in turn, place different demands on the productive, managerial, supportive, and other levels (cf. Hunt, 1970; Udy, 1965).

Thus a strong case has been made by several recent theories of organization for the recognition of the nature of the interaction between a social system and its environment, on the one hand, and its people on the other. By the use of the open-system model as a framework and conceptual language for describing organizations there is the possibility for a more thorough (but also more complex) understanding of organizational behavior:

> It is too soon to predict whether this approach will become a dominant guide to empirical work or whether "system" will become merely the latest "in" word among organizational psychologists [Quinn & Kahn, 1967, p. 461].

Contribution of Role Theory

In years past, the mind-body problem was a leading topic for philosophical discussion. For many, the tenability of the mind-body model depended on discovering the precise locus of the meeting of the two. The contemporary person-environment dilemma has prompted a similar search for a conceptual meeting point for social structural forces and individual personalistic forces. Programmatically, at least, probably the most satisfactory solution came about when Linton (1936) gave the notion of *role* a central place in the social sciences. Introduced to social psychology by Newcomb (1950), this concept has been explicated and advanced by such writers as Merton (1957), Shibutani (1961), Banton (1965), and Sarbin and Allen (1968). Until rather recently, however, the concept of role occupied only a secondary position among social science concepts, partly because earlier sociological and anthropological role notions tended to focus solely upon structural influences on personality. The advent of systems approaches to the study of large groups, however, gave the concept of role a new lease on life by balancing it with the personalistic forces that have traditionally been the concern of psychology.

Somewhere within the action system that is the functioning organization, structural and personal forces jointly draw a bead, as it were, on the individual, controlling his performances. And we have now reached the point of development in role concepts where, in their exemplary analysis of organizations, Katz and Kahn (1966) proposed role concepts as:

> the major means for linking the individual and organizational levels of research and theory; it is at once the building block of social systems and the summation of the requirements with which such systems confront their members as individuals [p. 197].

Highlighting distinctions between structure and person while concerned with their integration, Hunt (1967) pointed out that a social system may be viewed conceptually as an interlocking complex of positions (the structural aspect), or functional divisions of labor, which are to be populated by people who are expected to fulfill certain behavior requirements and possess certain personal attributes associated with each position. Since a social system is a system of interdependent activities, positions within a social structure exist in relationships with certain other positions. Thus, the "rights and duties" of a particular *focal position* exist with respect to a *counter-position*.

> This complementary contrast is, of course, a function basically of the complex patterns of behavior organized around these positions and embodying the relevant mutual expectations (the rights and duties) held by occupants of positions. *vis à vis* one another It is possible, therefore, to regard social process as an interaction of positions patterned in terms of these complementary expectations which are themselves called *roles* [p. 259].

Thus, role and its personalistic correlate, identity, represent the implications of social position incumbency and can be comprehensively described only with reference to other roles, which bear a complementary relation to the focal role.

We cannot explicate the ideas here, but we can say simply that since persons occupy multiple positions in life and are only partly involved in any single position they occupy, they have multiple identities that combine in various ways to affect their views and the enactments of their singular roles. And, whatever else may be involved, the modes of a man's participation in structured social intercourse will be reflected in his concept of himself and in the fabric of his personality.

The essential point, then, is not simply that role serves a linking function; more than that. In its modern forms, it exemplifies and operationalizes the merging of social and individual phenomena heretofore treated separatistically. Roles are undoubtedly social phenomena. However, they are not only external "demands," they are dynamic interactional processes carried out by individuals who color their performances personal. By way of reciprocity, however, through their identity implications roles become operationally integral to individual personality. Thus, roles more than link the individual and the social (or structural)—they unite them.

Role Conflict

Cyert and MacCrimmon (1968) made some of these same points, in their discussion of organizational processes, and accorded a prominent place to role concepts. In particular, they identified organizational roles as points of linkage and also as principal sources of conflict between individuals and organizations. In the study cited earlier, Hunt (1967) explicated five role varieties that are helpful in conceptualizing role differentiation. These focus attention on the interplay of patterns of expectation and performance within the role system. The importance of the distinctions he draws between aspects of role taking is that a lack of fit between any of them will result in some degree of role conflict, the effects of which have been the topic of two relatively recent major organizational studies and resulting theories.

Gross, Mason, and McEachern (1958) studied the perception and resolution of role conflict among school superintendents. They focused on both interrole conflicts (resulting from the occupancy of multiple positions) and intrarole conflicts (resulting from the perception of conflicting sent-roles[5] vis à vis a single focal position). On the basis of their data, they concluded that there were three major cognitive styles employed by superintendents in the resolution of perceived role conflict.

The first style they call the *moral* orientation. Individuals in this category give most weight to evaluating the legitimacy of the conflicting expectations or the *right* of others to hold such expectations. With minimal attention to the likelihood of sanctions for nonconformity, this style emphasizes conformity to the demand seen as most legitimate. The second type of orientation is called *expedient*. Priority here is given to a consideration of the sanctions others will bring to bear if the individual does not conform to their expectations. The resolution style in this case is to minimize negative sanctions involved in the role-conflict situation. The nature of the right of others to hold particular expectations is not nearly so important in this case as the perceived severity of the probable sanctions. This superintendent, therefore, will choose to conform to the expectations of those whose sanctions he fears most. *Moral expedient* is the third orientation toward role conflict resolution. Here no primacy is given to the dimensions of legitimacy or sanctions; rather, both are considered equally in choosing a course of resolution.

Among the numerous empirical findings of these authors was the high job satisfaction reported among superintendents who perceived agreement between the sent-role from the school board and their own role stereotypes. What is probably conceptually more important was the consistent finding of Gross et al. (1958) that the effects of role conflict on job satisfaction and expressions of worry among the superintendents were mediated by the characteristic anxiety level of the individual. Implicit in this finding is the notion that personality may be the result of an interaction of social structural conditions and personalistic dispositions. The difficulty in this particular study is that Gross and his colleagues dealt only with perceptions of role conflict and did not attempt to deal with the accuracy of perception. So the social structure is not truly represented in their research.

Clearly the most ambitious attempt to employ the role concept in organizational studies is that of Kahn et al. (1964). Borrowing Merton's (1957) formulation of a role set, these authors built a model that is specifically designed to integrate properties of a social structure with characteristics of people who populate it. They called their model a role episode and posited that it consists of a series of events. Specifically, role pressures emanate from members of the role set—the discrepancy between their own expectations and their evaluations of how well the focal person's behavior measures up. Members of the role set attempt to reduce this discrepancy by exerting pressures on the focal person to adjust his behavior to conform with their expectations. These pressures induce perceptual and cognitive experiences in the focal person that can lead to certain adjustive or maladjustive responses. The role-senders monitor any subsequent behavior of the focal person and reward him if they are satisfied with his adjustment or exert further pressures if they are not. This is a cyclic, ongoing process.

---

[5] Communications by persons intended to influence the role behavior of others (cf. Kahn, Wolfe, Quinn, Snoek, & Rosenthal, 1964).

Role conflict and ambiguity may thus occur in two ways. In the first case it may occur as a property of the social structure where, in fact, the sent-roles from role-set members exert pressures for various mutually exclusive behaviors on the part of the focal person (cf. also Hunt, 1967). Role ambiguity exists where the sent-role expectations lack sufficient clarity to be translated into behavior by the focal person. The second variety of role conflict occurs at the perceptual level and may or may not correspond with an actual state of affairs. Thus the focal person may not perceive conflict where it actually exists, or he may perceive it where it does not exist, or he may perceive it where it does indeed exist.

Furthermore, Kahn and his associates hypothesized that the degree of structural and perceived role conflict varies as a function of several enduring organizational and personal characteristics. For example, organizational positions vary in their vulnerability to conflicting expectations. Typically, the less structured and explicit the job requirement, the closer the position is to the boundary of the organization, and the larger the number of bosses, subordinates, and equals associated with a given position, the greater is the likelihood that role conflict (and the tensions and frustrations, which may occur as a result of its experience) will vary with such personalistic dispositions as needs for cognition, tolerance of ambiguity, neurotic anxiety, or need achievement.

Thus, these authors posited that the degree of role conflict and its relation to individual experience and reactions is highly conditional and depends on numerous organizational and personal attributes. Furthermore, variables such as job satisfaction and general mental health are anchored, via the role-set concept, to both the social structure and individual dispositions. These authors have further contributed to knowledge by operationalizing and testing their hypotheses—successfully.

The theories and data of the Kahn et al. (1964) study and the Gross et al. (1958) study have seriously challenged many old and static notions of organizational behavior. They have demonstrated crucial functions of personalistic dispositions as mediators between formal system requirements and individual outcomes. In other words, neither can stand alone as a comprehensive explanation of human behavior in organizations. Furthermore, they have provided evidence of the inadequacy of the personality versus organization hypothesis. While Gross et al. have shown that simply holding a high-ranking position does not render a man free from emotional problems, Kahn et al. have provided data comparing six hierarchical organizational levels and showing that tension induced by role strain increases with increasing rank. Since role conflict is linked with emotional reactions and mental health in general, it is apparently true that if the right measures are chosen, management positions may not be as rosy as the personality-organization theorists hold. Kahn et al. have clearly shown that the organizational locus of emotional maladjustment can occur anywhere in an organization and is not necessarily restricted to unskilled or low level workers. Employee reactions apparently depend on the nature of the organization, the nature of the job, and the nature of the person. It seems plain that Katzell's (1962) thesis warrants being taken seriously (and Quinn & Kahn's, 1967, as well). It is neither meaningful nor useful to promote normative, one-sided theories intended to account for all organizational situations. There is good evidence that organizational behavior is the outcome of a variety of highly conditional and highly contingent relationships and situations. Future theory will need to build on the foundations of those premises.

REFERENCES

Argyris, C. *Personality and organization.* New York: Harper, 1957.

Argyris, C. Understanding human behavior in organizations: One viewpoint. In M. Haire (Ed.), *Modern organization theory.* New York: Wiley, 1959.

Argyris, C. *Integrating the individual and the organization.* New York: Wiley, 1964.

Banton, M. *Roles.* New York: Basic Books, 1965.

Barker, R. G. On the nature of the environment. *Journal of Social Issues,* 1963, **19**, 17–38.

Bass, B. M. *Organizational psychology.* Boston: Allyn & Bacon, 1965.

Bell, D. *Work and its discontents.* Boston: Beacon Press, 1956.

Bennis, W. G. Organizational developments and the fate of bureaucracy. In L. L. Cummings & W. E. Scott (Eds.), *Organizational behavior and human performance.* Homewood, Ill.: Irwin-Dorsey, 1969.

Bennis, W. G. Organic populism. *Psychology Today,* 1970, **3**, 48–71.

Bennis, W. G., & Shepard, H. A theory of group development. *Human Relations,* 1965, **9**, 415–457.

Bennis, W. G., & Slater, P. E. *The temporary society.* New York: Harper & Row, 1968.

Blake, R. R., & Mouton, J. *Corporate excellence through grid organizational development.* Houston, Tex.: Gulf Corporation, 1968.

Blau, P. A formal theory of differentiation in organizations. *American Sociological Review,* 1970, **35**, 201–219.

Blauner, R. *Alienation and freedom: The factory worker and his industry.* Chicago: University of Chicago Press, 1964.

Buckley, W. (Ed.) *Modern systems research for the behavioral scientist.* Chicago: Aldine, 1968.

Caplow, T., & McGee, R. *The academic marketplace.* New York: Basic Books, 1958.

Carey, A. The Hawthorne studies: A radical criticism. *American Sociological Review,* 1967, **32**, 403–417.

Cleland, D. I., & King, W. R. *Systems analysis and project management.* New York: McGraw-Hill, 1968.

Coch, L., & French, J. R. P. Overcoming resistance to change. *Human Relations,* 1948, **1**, 512–533.

Crozier, M. *The bureaucratic phenomenon.* Chicago: University of Chicago Press, 1964.

Cyert, R. M., & MacCrimmon, K. R. Organizations. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology.* Vol. I. (2nd ed.) Boston: Addison-Wesley, 1968.

Cyert, R. M., & March, J. G. *A behavioral theory of the firm.* Englewood Cliffs, N. J.: Prentice-Hall, 1963.

Dawis, R. V., Lofquist, L. H., & Weiss, D. J. A theory of work adjustment. *Minnesota Studies in Vocational Rehabilitation,* 1968, **23**, Bulletin 47.

Durkheim, E. *De la division du travail social.* [Division of labor in society] (Translation, G. Simpson, New York: Macmillan, 1933, 1947) Paris: F. Alcan, 1902.

Festinger, L. Informal social communication. *Psychological Review,* 1950, **57**, 271–282.

Fiedler, F. E. Engineer the job to fit the manager. *Harvard Business Review,* 1965, **43**, 115–122.

French, J. R. P., Israel, J., & Aas, D. An experiment on participation in a Norwegian factory. *Human Relations,* 1960, **13**, 3–19.

Galbraith, J. K. *The new industrial state.* Boston: Houghton Mifflin, 1967.

Gellerman, S. W. *Motivation and productivity.* New York: American Management Association, 1963.

Gilbreth, F. B. *Motion study.* New York: Van Nostrand, 1911.

Glans, T. B., Grad, B., Holstein, D., Meyers, W. E., & Schmidt, R. N. *Management systems.* New York: Holt, Rinehart & Winston, 1968.

Gordon, M. M. *Assimilation in American life: The role of race, religion, and national origins.* New York: Oxford, 1964.

Gouldner, A. W. *Patterns of industrial bureaucracy.* New York: Free Press, 1954.

Gross, N., Mason, W. S., & McEachern, A. W. *Explorations in role analysis.* New York: Wiley, 1958.

Gulick, L., & Urwick, L. (Eds.) *Papers on the science of administration.* New York: Institute of Public Administration, 1937.

Haberstroh, C. J. Organization design and systems analysis. In J. March (Ed.), *Handbook of organizations.* Chicago: Rand McNally, 1965.

Hackman, R. *The motivated working adult.* New York: American Management Association, 1969.

Haire, M. Philosophy of organizations. In D. M. Bowerman & F. M. Fillerup (Eds.), *Management: Organization and planning.* New York: McGraw-Hill, 1963.

Hall, R. The concept of bureaucracy: An empirical assessment. *American Journal of Sociology,* 1963, **69**, 32–40.

Harvey, E. Technology and the structure of organizations. *American Sociological Review,* 1968, **33**, 247–259.

Herzberg, F., Mausner, B., & Snyderman, B. *The motivation to work.* (2nd ed.) New York: Wiley, 1959.

Hickson, D. J., Pugh, D. S., & Pheysey, D. C. Operations technology and organization structure. *Administrative Science Quarterly,* 1969, **17**, 378–397.

Hollander, E. P., & Hunt, R. G. (Eds.) *Current perspectives in social psychology.* New York: Oxford, 1967.

Homans, G. C. *The human group.* New York: Harcourt, Brace & World, 1950.

Hunt, R. G. Role and role conflict. In E. P. Hollander & R. G. Hunt (Eds.), *Current perspectives in social psychology*. (2nd ed.) New York: Oxford, 1967.

Hunt, R. G. Review of E. J. Miller and A. K. Rice, *Systems of organization. Administrative Science Quarterly*, 1968, **13**, 360–362.

Hunt, R. G. An essay on the profit motive. *Defense Management Journal*, 1969, **5**, 6–11. (a)

Hunt, R. G. Review of R. C. Hackman, *The Motivated Working Adult. Administrative Science Quarterly*, 1969, **14**, 614–615. (b)

Hunt, R. G. Technology and organization. *Academy of Management Journal*, 1970, **13**, 235–253.

Indik, B. P. Some effects of organization size on member attitudes and behavior. *Human Relations*, 1963, **16**, 369–384.

Indik, B. P. Organization size and member participation: Some empirical tests of alternative explanations. *Human Relations*, 1965, **18**, 339–350.

Jasinski, F. J. Use and misuses of efficiency controls. *Harvard Business Review*, 1956, **34**, 105–112.

Kahn, R. L., Wolfe, D. M., Quinn, R. P., Snoek, J. D., & Rosenthal, R. A. *Organizational stress: Studies in role conflict and ambiguity*. New York: Wiley, 1964.

Katz, D. The motivational basis of organizational behavior. *Behavioral Science*, 1964, **9**, 131–146.

Katz, D., & Kahn, R. L. *The social psychology of organizations*. New York: Wiley, 1966.

Katz, F. E. Explaining informal work groups in complex organizations: The case for autonomy in structure. *Administrative Science Quarterly*, 1965, **10**, 204–221.

Katzell, R. A. Contrasting systems of work organization. *American Psychologist*, 1962, **17**, 102–108.

Katzell, R. A., Barrett, R. S., & Parker, T. C. Job satisfaction, job performance, and situational characteristics. *Journal of Applied Psychology*, 1961, **45**, 65–72.

Koontz, H. Making sense of management theory. *Harvard Business Review*, 1962, **40**, 25.

Leavitt, H. J. Unhuman organizations. *Harvard Business Review*, 1962, **40**, 90–98.

Levinson, H., Price, C. R., Munden, K. J., Mandel, H. J., & Solley, C. M. *Men, management and mental health*. Cambridge: Harvard University Press, 1962.

Lewin, K. *Field theory in social science*. New York: Harper, 1951.

Lewin, K. Group decision and social change. In E. Maccoby, T. Newcomb, & E. Hartley (Eds.), *Readings in social psychology*. (3rd ed.) New York: Holt, Rinehart & Winston, 1958.

Likert, R. *New patterns of management*. New York: McGraw-Hill, 1961.

Likert, R. *The human organization*. New York: McGraw-Hill, 1967.

Linton, R. *The study of man*. New York: Appleton-Century, 1936.

Maier, N. *Psychology in industry*. Boston: Houghton Mifflin, 1955.

Mann, F. C., & Hoffman, L. R. *Automation and the worker: A study of social change in power plants*. New York: Holt, 1960.

March, J. G., & Simon, H. A. *Organizations*. New York: Wiley, 1958.

Marx, K. *Capital*. London: Allen & Unwin, 1887. (Translation reproduced and supplemented, 1943.)

Marx, K. *Selected writings in sociology and social philosophy*. (Edited and Translated by T. B. Bottomore & M. Rubel) New York: McGraw-Hill, 1964.

Maslow, A. H. *Motivation and personality*. New York: Harper, 1954.

Maslow, A. H. *Eupsychian management*. Homewood, Ill.: Irwin-Dorsey, 1965.

Mayo, E. *The human problems of industrial civilization*. New York: Macmillan, 1933.

McGregor, D. *The human side of enterprise*. New York: McGraw-Hill, 1960.

Mechanic, D. Some considerations in the methodology of organizational studies. In H. Leavitt (Ed.), *The social science of organizations*. Englewood Cliffs, N. J.: Prentice-Hall, 1963.

Merton, R. K. *Social theory and social structure*. (Rev. ed.) New York: Free Press, 1957.

Miller, E. J., & Rice, A. K. *Systems of organization: The control of task and sentient boundaries*. London: Tavistock, 1967.

Miller, J. G. *Living systems*. New York: Wiley, 1968.

Morane, J. H. *A sociology of human systems*. New York: Appleton-Century-Crofts, 1967.

Mouzelis, N. P. *Organization and bureaucracy*. Chicago: Aldine, 1967.

Neff, W. S. *Work and human behavior*. New York: Atherton, 1968.

Newcomb, T. M. *Social psychology*. New York: Dryden, 1950.

Newcomb, T. M. Individual systems of orientation. In S. Koch (Ed.), *Psychology: A study of a science*. Vol. 3. New York: McGraw-Hill, 1959.

Opsahl, R. L., & Dunnette, M. D. The role of financial compensation in industrial motivation. *Psychological Bulletin*, 1966, **66**, 94–118.

Parsons, T. *Structure and process in modern societies*. New York: Free Press, 1960.

Perrow, C. A framework for the comparative analysis of organizations. *American Sociological Review*, 1967, **32**, 194–209.

Porter, L. W., & Lawler, E. E. Properties of organization structure in relation to job attitudes and job behavior. *Psychological Bulletin*, 1965, **64**, 23–51.

Porter, L. W., & Lawler, E. E. *Managerial attitudes and performance*. Homewood, Ill.: Irwin-Dorsey, 1968.

Pugh, D. S. Modern organization theory: A psychological and sociological study. *Psychological Bulletin*, 1966, **66**, 235–251.

Quinn, A. P., & Kahn, R. L. Organizational psychology. *Annual Review of Psychology*, 1967, **18**, 437–466.

Rice, A. K. *The enterprise and its environment.* London: Tavistock, 1963.

Roethlisberger, F. J., & Dickson, W. J. *Management and the worker.* Cambridge: Harvard University Press, 1939.

Ronan, W. W. Individual and situational variables relating to job satisfaction. *Journal of Applied Psychology Monograph,* 1970, **54**, No. 1, Part 2.

Roy, D. F. Banana time: Job satisfaction and informal interaction. *Human Organization,* 1960, **18**, 158–168.

Rubin, I., & Hunt, R. G. *Some aspects of managerial control in interpenetrating systems: The case of government-industry relations.* (Tech. Rep. No. 7) State University of New York at Buffalo: Grant NGR 33-015-061, National Aeronautics and Space Administration, July 1969.

Sarbin, T. R., & Allen, V. L. Role theory. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology.* Vol. 1. (2nd ed.) Boston: Addison-Wesley, 1968.

Sayles, L., & Strauss, G. *Human behavior in organizations.* Englewood Cliffs, N. J.: Prentice-Hall, 1966.

Schein, E. H. *Organizational psychology.* Englewood Cliffs, N. J.: Prentice-Hall, 1965.

Schein, E. H., & Bennis, W. G. *Personal and organizational change through group methods: The laboratory approach.* New York: Wiley, 1965.

Schlesinger, A. *A thousand days.* Boston: Houghton Mifflin, 1965.

Scott, W. G. Organization theory: An overview and an appraisal. *Academy of Management Journal,* 1961, **4**, 7–27.

Scott, W. R. Theory of organizations. In R. E. L. Farris (Ed.), *Handbook of modern sociology.* Chicago: Rand McNally, 1964.

Shibutani, T. *Society and personality.* Englewood Cliffs, N. J.: Prentice-Hall, 1961.

Simon, H. *Administrative behavior.* (2nd ed.) New York: Macmillan, 1957.

Simon, H. *The shape of automation for men and management.* New York: Harper & Row, 1965.

Smith, P. C., & Cranny, C. J. Psychology of men at work. *Annual Review of Psychology,* 1968, **19**, 467–496.

Sommer, R. Small group ecology. *Psychological Bulletin,* 1967, **67**, 145–152.

Sommer, R. *Personal space.* Englewood Cliffs, N. J.: Prentice-Hall, 1968.

Strauss, G. Some notes on power equalization. In H. Leavitt (Ed.), *The social science of organizations.* Englewood Cliffs, N. J.: Prentice-Hall, 1963.

Strauss, G. Human relations—1968 style. *Industrial Relations,* 1968, **7**, 262–272.

Tannenbaum, A. S. *Social psychology of the work organization.* Belmont, Calif.: Wadsworth, 1966.

Tannenbaum, A. S. *Control in organizations.* New York: McGraw-Hill, 1968.

Taylor, F. W. *Scientific management.* New York: Harper & Row, 1911.

Thompson, J. D. *Organization in action.* New York: McGraw-Hill, 1967.

Thompson, V. *Modern organization.* New York: Knopf, 1961.

Trist, E. L., Higgin, G. W., Murray, H., & Pollack, A. B. *Organizational choice.* London: Tavistock, 1963.

Udy, S. H. The comparative analysis of organizations. In J. March (Ed.), *Handbook of organizations.* Chicago: Rand McNally, 1965.

von Bertalanffy, L. The theory of open systems in physics and biology. *Science,* 1950, **111**, 23–38.

Vroom, V. H. *Work and motivation.* New York: Wiley, 1964.

Vroom, V. H., & Mann, F. Leader authoritarianism and employee attitudes. *Personnel Psychology,* 1960, **13**, 125–140.

Walker, C. R. Life in the automatic factory. *Harvard Business Review,* 1958, **36**, 111–119.

Weber, M. *Theory of social and economic organization.* (Trans. by A. M. Henderson & T. Parsons of Pt. I, *Wirtschaft und Gesellschaft*) New York: Oxford, 1947. (2nd ed., Free Press, 1964.)

Whyte, W. F. *Organziation behavior: Theory and application.* Homewood, Ill.: Irwin-Dorsey, 1969.

Woodward, J. *Industrial organization: Theory and practice.* London: Oxford University Press, 1965.

(Received March 31, 1970)

Psychological Bulletin
1971, Vol. 76, No. 4, 295–310

PROBLEMS WITH THE USE OF "MODERATOR" VARIABLES

SHELDON ZEDECK[1]

University of California, Berkeley

Alternative concepts of "moderator" variables are defined and categorized in terms of whether they lead to differential validity, or lead to differential predictability, or involve moderated regression techniques. The dissimilarity of results provided by the three approaches is discussed in terms of the type of variable (qualitative versus quantitative) examined as a moderator and in terms of statistical and methodological problems. The major problems of the operation and identification of moderator variables also are discussed. Finally, utility of the moderator variable concept is examined.

In 1928, Hull noted that the upper limit of validity coefficients for predictors of criteria was of the order of .50. In 1966, Ghiselli reviewed unpublished and published validity research and found that a validity coefficient for a predictor rarely surpassed .50. The lack of improvement in prediction between the two reports generated modifications of the classic or traditional validation model.

The classic validation model provides a simple index of the relationship between predictor and criterion and ignores factors intervening between behaviors on the two variables. As a result, newer strategies for validation of predictors have been suggested. In particular, Guetzkow and Forehand (1961) suggested a prediction model to be used in selection of executives that requires the analysis of multiple predictors and multiple criteria as modified by multiple behavior types and situations. Dunnette (1963) extended the strategy and suggested the analysis of the complex interactions that may occur between the predictors and criteria; that is, different groups or types of individuals, different behaviors on the job, and different situations. This modified validation model allows for (*a*) predictors to be differentially valid for different groups of individuals, (*b*) similar job behaviors to be predictable by different patterns of interactions between predictors and individuals, and (*c*) similar job behaviors to have differential organizational consequences depending on the situation.

Considerable research has been devoted to the identification of homogeneous sets of individuals as a result of the modification of the classic validation model. These particular research endeavors generally have been classified as "moderator" variable research. The existing moderator techniques are not entirely independent and achieve the goal of identifying homogeneous sets of individuals either through psychometric manipulations or through classification procedures. However, analyses of the same set of data by the various techniques have produced differential results. Hence, the purpose of the present study is to differentiate the various moderator variable techniques and to present the problems with the use of moderator variables. In addition, the efficiency and potential utility of moderator variable approaches are discussed.

MODERATOR VARIABLES AND PROCEDURES

The term moderator variable is used to describe the notion of "population control variable" (Gaylord & Carroll, 1948), "subgrouping variable" (Frederiksen & Melville, 1954), "referrent variable" (Toops, 1959), "predictability variable" (Ghiselli, 1956), "modifier variable" (Grooms & Endler, 1960), and "homologizer variable" (Johnson[2]). In addition, Saunders' (1955) original definition of the term moderator variable has been generalized by Banas[3] to include qualitative as well as quantitative variables. The procedures used for the above concepts are different and, consequently, often lead to dissimilar results. For purposes of illustrating the techniques, the procedures are categorized in terms of whether they lead to differential validity (i.e., subgroups formed by fractionation on the moderator scores demonstrate different validity coefficients), lead to differential predictability (i.e., subgroups formed on the basis of the predictability of the individuals demonstrate different validity coefficients), or involve moderated regression techniques (i.e., one moderated regression equation which is developed for the total group provides improved prediction over that of ordinary multiple regression). The latter categorization centers on the moderator variable as a psychometric characteristic, whereas the former categorizations center on the moderator variable as descriptions of relationships between two variables (either in terms of predictability or in terms of possession of a characteristic).

---

[1] Requests for reprints should be sent to Sheldon Zedeck, Department of Psychology, University of California, Berkeley, California 94720.

[2] Johnson, C. D. The population control variable or moderator variable in personnel research. In Tri-Service Conference on Selection Research. Washington, D. C.: Office of Naval Research, 1960, 125-134.

Differential Validity

These techniques examine the influence of a third variable, qualitative or quantitative, on a predictor-criterion relationship. The third variable is used as a basis for splitting the total group into subgroups. That is, r_{XY} for Group A is significantly different from r_{XY} for Group $\bar{A}$ (not A), where A is a qualitative or quantitative descriptive variable.

Subgrouping variables. A subgrouping variable is the basis for splitting a sample into subgroups with the result that the validity coefficient of the predictor variable for one subgroup is significantly different from the coefficient of the other subgroup. The technique has been investigated in studies pertaining to education as early as 1926 when Scates[4] found that grades of college students graduated from certain high schools varied as a function of location and size of town. Similar results were obtained by Wagner and Strabel (1935) who subgrouped based on sex, residency, high school, age, degree sought, and other population variables. In addition, sex (Abelson, 1952) and degree of adjustment (Hoyt & Norman, 1954) have been found to provide differential validities for predictors of success in academics.

In a frequently cited study of subgroup analysis, Frederiksen and Melville (1954) hypothesized that the usefulness of an interest test is improved by identifying subgroups of male engineering students for whom the test was especially appropriate as a predictor; specifically, the usefulness of interest measures is limited to a relatively noncompulsive group. A measure of reading speed and an unspeeded vocabulary test were used to identify compulsive and noncompulsive subjects. In 5 of 10 interest scales on the Strong Vocational Interest Blank, correlations between interest scores and grade point average were significantly higher for the noncompulsive group than for the compulsive group. A replication by Frederiksen and Gilbert (1960) yielded similar results but only for the occupational keys most logically related to engineering. Stricker (1966) examined the generality of the findings with liberal arts and science students of both sexes. Results indicated that the compulsivity scales were not indicating subgroups with different coefficients.

A considerable number of subgrouping analyses have been reported since the Frederiksen and Melville (1954) study. Variables that have been examined for subgrouping characteristics include demographic factors and biographical information blanks (Baker;[5] Leitch;[6] Medvedeff;[7] Rock;[8] Tesser, Starry,

---

[3] Banas, P. A. An investigation of transsituational moderators. Unpublished doctoral dissertation, University of Minnesota, 1964.

[4] Scates, D. E. Selective admission and retention of college students at the University of Chicago. Unpublished doctoral dissertation, University of Chicago, 1926.

[5] Baker, B. L. The use of biographic factors to moderate prediction and predict salary level. Unpublished doctoral dissertation, Purdue University, 1967.

[6] Leitch, D. W. An initial investigation of a multidimensional validation model and the derivation of homogeneous subsets. Unpublished doctoral dissertation, Purdue University, 1965.

[7] Medvedeff, E. The utility of sub-grouping analysis in the prediction of sales success. Unpublished doctoral dissertation, Purdue University, 1964.

[8] Rock, D. A. Improving the prediction of academic achievement by population moderators. Unpublished dissertation, Purdue University, 1965.

& Channey, 1967), self-esteem (Korman, 1967, 1969), self-report personality variables with respect to motivation (Brown[9]), motivation in risk-taking situations (Kogan & Wallach, 1964), problem-solving styles (French, 1965), knowledge and aptitude tests (Prediger, 1966, 1969; Steinemann, 1964), superior's ranking of ability (Lawler, 1966), intelligence levels (Goodstein & Heilbrun, 1962), motor skills (Kipnis, 1962), job characteristics (Peterson, 1964), job and organizational factors (Carroll & Tosi, 1969), occupation (Doll & Gunderson, 1969; Dunnette & Kirchner, 1960), and job satisfaction (Dawis, Weiss, Lofquist, & Betz, 1967).

Referrent variable. Toops (1959) proposed that the meaning of an individual's response is relative to the homogeneous group, or "ulstrith" group, to which he belongs. In addition, the responses are relative to referrent variables (status and environmental variables) which modify the meaning or causal potency of other variables. The weight of an independent variable is not constant, but rather is a mathematical function of the referrent variable. The implication of interactions suggests the use of moderated regression equations. Referrent variables, however, are mentioned in the differential validity section because of the emphasis by Toops on ulstriths and classification of profiles. The status and environmental variables (education, age, etc.) examined as subgrouping variables also can be termed "referrent variables."

Modifier variable. Grooms and Endler (1960) formed three anxiety subgroups (high, medium, and low) and examined the differences between them with respect to the relationship between an aptitude predictor and grade point criterion. Grooms and Endler termed anxiety a modifier variable because differential validities were found between anxiety subgroups. A modifier variable is an independent variable which, when dichotomized or trichotomized, leads to differential subgroup relationships between a predictor variable and a criterion variable. Again, there is no difference between anxiety as a referrent variable or anxiety as a subgrouping variable.

Differential Predictability

Differential predictability variables also identify groups with differing validities. In this case, however, the subgroups are formed on the basis of predictability; that is, the variable identifies subgroups of individuals who differ in predictability with respect to a particular predictor-criterion relationship. Thus, r_{XY} for relatively predictable individuals (small deviations from the regression line) is significantly greater than r_{XY} for nonpredictable individuals.

Predictability variable. A predictability variable is a variable which is correlated with an absolute difference score between standardized predictor (Z_p) and standardized criterion (Z_c) scores (Ghiselli, 1956). The predictability variable thus identifies those individuals in a sample who deviate from a regression line and for whom the predictor is inappropriate.

Ghiselli's (1956) technique involves obtaining absolute difference scores, $|D|$, between Z_c and Z_p. Algebraic differences are not used since the model is concerned with accuracy of prediction, and overprediction and underprediction of the same degree are considered equal errors. The magnitude of $|D|$ serves as an index of predictability; the smaller the $|D|$, the better the relation between criterion and predictor. Correlates of the $|D|$ scores are subsequently identified and used as "predictors of predictability." A high positive correlation indicates that the validity of the predictor should be higher for those scoring low on the "predictor of predictability" than for the entire group. The difference between the validity coefficients for the lowest (e.g., $\frac{1}{3}$) scorers and the highest ($\frac{2}{3}$) scorers on the predictability test is then tested for significance. A significant difference indicates that the predictor should be used only for the low scorers and a different predictor should be used for the remaining subjects. Johnson (see Footnote 2) refers to a variable with the above-mentioned properties as an *homologizer variable* (a variable used to group scores of equal credibility or information value).

---

[9] Brown, O., Jr. The utility of self-report variables as predictive moderators. Paper presented at the meeting of the Midwestern Psychological Association, Chicago, May 1968.

Ghiselli (1956) demonstrated the potential usefulness of the technique in a study involving prediction of job proficiency of taxicab drivers. An occupational inventory identified those individuals for whom a tapping-and-dotting test was a good predictor of job proficiency.

In essence, two rejection steps of individuals are involved. First, applicants are rejected based on scores on the predictability variable. Second, individuals are rejected based on scores on the predictor test. The optimal proportion eliminated in each screening is a function of the magnitude of the correlation between the tests, the criterion, and the difference scores. Best results in the taxicab study (Ghiselli, 1956) were obtained when a larger proportion was eliminated in the first than in the second screening.

Ghiselli (1960b) also demonstrated that scales of an inventory used as a predictor could serve as its own predictability test. In this technique, $|D|$ scores are obtained for each individual in an experimental group. The original items of the predictor are then reanalyzed to determine which items discriminate those individuals with large $|D|$ scores from those with small $|D|$ scores. Low scores on the resulting predictability scales imply small differences between Z_c and Z_p and high scores imply large differences. In the Ghiselli study, the predictability scales were applied to a cross-validation group for which the hypothesized relationships held. Essentially, this procedure involves a refinement of the predictor and not a moderator study, which implies the influence of a third variable on two others.

Ghiselli (1963) found that the technique resulted in the identification of moderators that are highly situation-specific. Banas (see Footnote 3), however, found that *transsituational moderators* (moderators that function to identify more and less predictable subgroups when applied in different situations) obtained by Ghiselli's technique do exist in the prediction of performance of three classes of workers in a rehabilitation center.

In another application of the technique, Ghiselli and Sanders (1967) developed a predictability test from a self-analysis inventory that differentiated two subgroups with different patterns of heteroscedasticity.

Blood[10] found demographic variables that differentiated a job level–job satisfaction relationship. In this case, the Ghiselli (1956) approach was modified in that a multiple correlation was examined to determine the relationship between $|D|$ and several demographic variables. As a result, several variables operated simultaneously as the predictability variable. The demographic variables included continuous and discrete variables.

Ghiselli (1960a) extended the predictability variable concept and demonstrated that it is possible to differentiate which of two tests gives better prediction for a given individual. For two tests, Test 1 and Test 2, $|D_1|$ and $|D_2|$ are obtained, respectively. For all individuals for whom $|D_1| < |D_2|$, scores on Test 1 predict criterion scores more accurately than do scores on Test 2; for all individuals for whom $|D_1| > |D_2|$, scores on Test 2 predict criterion scores more accurately than do scores on Test 1. Consequently, we obtain $|D_2| - |D_1|$ for each individual and thereby have a set of scores such that positive values indicate individuals for whom Test 1 provides better prediction, and negative values indicate individuals for whom Test 2 provides better prediction. Correlates of the difference between $|D_2|$ and $|D_1|$ are sought; the correlates are predictability variables. Critical scores on the predictability variable indicate whether Test 1 or Test 2 should be used for the individual.

Ghiselli (1960a) successfully utilized the above-described procedure in three studies using personality and interest inventories as the predictability variables. Richardson (1965), however, failed in three attempts to improve prediction of college grade point average with items from the California Psychological Inventory or Minnesota Multiphasic Personality Inventory used as predictability variables. Richardson suggested that the failure was due to the complexity of the factors involved in grade point average.

---

[10] Blood, M. R. Job level and job satisfaction: A contingent relationship. Paper presented at the meeting of the Midwestern Psychological Association, Chicago, May 1967.

The possibility exists, however, that more stable results would be obtained if the predictability variable were correlated with each of the $|D|$ scores independently, rather than correlated with the difference between $|D_2|$ and $|D_1|$. It is possible to obtain the predictability relationship when, for example, D_1 is correlated with the predictability variable (and, therefore, appropriate for some of the individuals), and D_2 is uncorrelated with the predictability variable (and, therefore, inappropriate for all individuals). Furthermore, it is possible to find correlates of the difference between $|D_2|$ and $|D_1|$ when both predictors are relatively inefficient. These possibilities can occur and therefore provide misleading information, when both D_1 and D_2 have high absolute values, for instance, 10 and 9, respectively. The difference between them is small and negative and would indicate that Test 2 is more appropriate. However, the high absolute values of each indicate that both tests are inappropriate.

Quadrant analysis. Marks (1964) suggested that researchers direct their attention to situations in which indicated relationships fail; that is, to study off-quadrant relationships. In following Marks' suggestion, Hobert and Dunnette (1967) claimed that there is practical value in identifying overpredicted ($Z_p > Z_c$) and underpredicted ($Z_p < Z_c$) individuals as separate groups rather than treating them as an aggregate of unpredictables. Underpredicted and overpredicted individuals are identified by examining a scatter diagram depicting the bivariate distribution of a predictor test and the corresponding criterion scores. Both the predictor and criterion scores are divided at the median to yield the following classifications of individuals: High hits (high predictor-high criterion), low hits (low predictor-low criterion), overpredicted (high predictor-low criterion), and underpredicted (low predictor-high criterion).

Hobert and Dunnette (1967) developed two moderators: one for the underpredicted and low hits, and the other for the high hits and overpredicted. Moderators were developed through an analysis of the items used to form the predictor composite; that is, those items which discriminated significantly between criterion score levels within a given predictor score level (e.g., between the low hits and underpredicted groups) were used. Hobert and Dunnette claimed that two moderators were operating, one for each level of predictor score, which enhanced the prediction of managerial effectiveness.

Abrahams (1969), in a comment on quadrant analysis, stated that mean differences on the predictor composite (e.g., low hits and underpredicted) are ignored. Since the moderator is composed of items from the predictor composite, differences found in item analyses reflect to some degree these predictor mean differences. The developed moderator variable may be predictive only of differences on the predictor variable itself. At best, the components of the moderator variable would reflect a combination of differences, that is, actual differences between hits and misses and spurious differences due to predictor score differences. The items and variables showing such differences are related to the predictor, and it is doubtful that such a moderator has value beyond that already shown in the predictor-criterion relationship. In essence, as with Ghiselli's (1960b) use of items of the predictor of predictability test, the predictor items' weights are changed.

Similarly, McNemar (1969) demonstrated that when the items and scales of the original predictor composite are used in the moderator without the restriction that the moderator be uncorrelated with the predictor, the predictor itself would be a moderator. The demonstration revealed increased predictive efficiency, increased hit rate, and decreased percentage of overlap between high- and low-criterion groups. However, though the validity coefficient of the predictor increases, the error of estimate remains unchanged when cases are screened out from the central portion of the distribution on the predictor variable. In addition, when the gain in hit rate is based on the total sample rather than on the sample after screening, the actual result is a loss in hit rate.

Applications of quadrant analysis have

demonstrated little success. Blood (see Footnote 10) used quadrant analysis with job level–job satisfaction data. No moderators were found. Berdie (1969b) examined the effects of intraindividual variability indexes on a predictor-criterion relationship. Again, there were few consistent or meaningful trends. In both studies, potential moderators were independent of the predictor variables.

Multipredictable group validation. The concept of quadrant analysis or any model involving off-quadrant consideration is a logical extension of Ghiselli's (1956) technique. Emphasis on underpredicted and overpredicted individuals, as opposed to combining these individuals into an unpredictable group, has practical implications. Maintaining differences among unpredictable groups provides enhancement of the understanding of the psychological composition of groups and, in prediction, aids in the identification of false positives and false negatives. Hence, multipredictable group validation (Zedeck[11]) is another extension of Ghiselli's (1956) technique. Multipredictable group validation involves algebraic differences (considers underpredicted, overpredicted, and predicted individuals), $D = Z_c - Z_p$, and deviations from the regression line. The underpredicted group is composed of individuals with D greater than some a priori value, for instance, $> + 1$. The overpredicted group is composed of individuals with $D < - 1$. The predicted group is composed of individuals with $-1 \leq D \leq +1$. Size of the predicted region and number of levels of underpredicted and overpredicted groups could vary depending on the situation. In one study (Zedeck, see Footnote 11), using the multipredictable group validation model in an educational situation, results did not hold up in cross-validation.

Intraindividual variability. Differences in intraindividual variability are related to predictability. Berdie (1961) used an index of intraindividual variability as a potential moderator variable. The basic notion of the concept implies that intraindividual variability indexes determine subgroups for whom more efficient prediction can be made. The specific hypothesis was that students whose variability of performance (scores on 10 mathematics subtests) was greater while taking a test used for predictive purposes would be less predictable than students whose behavior was less variable. Again, the procedure is similar to Ghiselli's (1960b) use of a predictor as its own predictability test. The results of Berdie's (1961) study, however, were not statistically significant.

A more recent study by Berdie (1969b) investigated the relationship between intraindividual variability and predictability. The evidence suggested that the relationship is complex and consideration must be given to the individual's level of performance on the tasks from which the variability indexes are developed, the level of performance on the predictor and criterion, and the method used in determining accuracy of prediction.

Berdie (1969a) further studied the extension of the concept and found that the variability of an individual on one task is somewhat related to his variability on certain other tasks. Berdie, however, cautioned that the tasks on which conclusions are made must be specified. In a sense, the results argue for the existence of transsituational variables.

Moderated Regression Techniques

Moderated regression techniques involve prediction equations for the total sample and do not require explicit subgrouping or differentiating of individuals. The moderated regression equation contains a higher-order term such as an xz cross-product or x^2. A test is made to determine whether the higher-order term in the moderated regression equation contributes to prediction beyond that of the ordinary linear regression equation (without the additional higher-order term).

Population control variable. Gaylord and Carroll (1948) defined a population control variable as a variable that identifies subpopulations in which the application of a multiple regression equation, optimal for the entire population, is inappropriate. A multiple regression equation is developed that includes products and cross-products of the variables in a multivariate analysis. Any variable in the analysis may have population control as well as predictor effects.

---

[11] Zedeck, S. Identification of moderator variables by discriminant analysis in a multi-predictable group validation model. Unpublished doctoral dissertation, Bowling Green State University, 1969.

Johnson (see Footnote 2) suggested that the two variables used by Gaylord and Carroll (1948), military status and degree of rating contact with the ratee, be referred to as *population control executive variables*. Johnson claimed that the variables did not control extraneous variance but, instead, served to direct which additional variables should be used in a second prediction stage.

Moderator variable. Saunders (1955, 1956) coined the term moderator variable to indicate *continuous* variables that influence the predictive effectiveness of the predictor variable. A multivariate curvilinear regression equation involving cross-products is used in which the beta weights, instead of being constant, are linear functions of moderator variables.

The moderator variable is a means of maintaining the integrity of the total population while still maintaining a statistical control on each individual's membership in one of a continuous, infinite series of subpopulations defined by his score on the moderator (Saunders, 1955). Saunders (1956) treated the cross-product of the moderator and predictor scores as a new predictor in any standard multiple regression technique. The regression equation is $y = \bar{y} + \Sigma a_i x_i + \Sigma b_j z_j + \Sigma c_{ij} x_i z_j$, where a, b, and c are weights, x_i the predictors, and z_j the moderators.

Saunders (1956) reanalyzed the Frederiksen and Melville (1954) data using the moderated regression technique. Moderated regression resulted in increases in predictive validity over the multiple correlation method in 5 of 10 comparisons. However, in no case did moderated regression provide significant improvement over the coefficient for the noncompulsive group.

Kirkpatrick, Ewen, Barrett, and Katzell (1968) investigated "race" and "cultural deprivation" (measured by factor scores) as moderators of success in various occupations. In general, the data did not strongly support their hypotheses that tests are differentially valid for different ethnic groups.

Cleary (1966) presented a generalized moderated regression technique that allows individual differences to emerge without the constraints of a priori conceptions. One set of weights is derived for each person on a dimension, and a second set for each predictor. The two sets of weights are combined into a composite weight for each dimension in the combined predictor-criterion matrix. This individual differences model showed some improvement in prediction over the regular multiple regression technique.

Comparative Efficiency of Moderator Variable Procedures

Common to the alternative definitions and procedures is the notion of dividing a heterogeneous population into homogeneous subgroups on the basis of the variables affecting the relationship between predictor and criterion. Thus, interaction among variables is implied. Another implication is that homogeneous subgroups, showing differential patterns of validity, may be isolated (Banas, see Footnote 3). Consideration of these implications, in addition to the frequent interchange of terms, has caused Banas to define a general moderator variable as a qualitative or quantitative variable that improves the usefulness of a predictor by isolating subgroups of individuals for whom a predictor or set of regression weights are especially appropriate. Consequently, it is implied that the same moderator variable should be identified regardless of the procedure used. Several studies have examined the efficiency of the procedures relative to each other and to linear regression.

Saunders (1956) compared moderated regression to subgroup analysis. Results showed that moderated regression was an improvement over linear regression. As previously mentioned, however, the moderated regression did not provide an improvement over the validity coefficient for the noncompulsive subgroup. Parrish[12] compared the predictive efficiency of moderated regression with that of linear multiple regression. Aptitude, interest, and personality scales were used as predictors, and combat efficiency ratings were used as a criterion. Results indicated that moderated regression was not more effective than linear multiple regression.

---

[12] Parrish, J. A. A study of two non-linear methods of combining predictor tests. (PRB Technical Note 103) Washington, D. C.: Personnel Research Branch, AGO, Department of the Army, 1959, 1–21.

Johnson (see Footnote 2) compared moderated regression techniques with simple regression and with moderated regression within subgroups. Results showed that ordinary subgrouping was more effective than moderated regression for the total group or for the subgroups. In addition, the difference in effectiveness between a linear model and the moderator model is less for highly selected groups (selected on the basis of a moderator variable) than for the total group.

Zedeck, Cranny, Vale, and Smith[13] compared moderated regression, multiple regression, Ghiselli's (1956) procedure, and subgroup analysis. Variables effective as moderators in subgroup analysis and prediction of predictability were not effective in moderated or multiple regression.

Blood (see Footnote 10) analyzed his data with Ghiselli's (1956) technique and with quadrant analysis (Hobert & Dunnette, 1967). Results showed that Ghiselli's (1956) procedure was more successful than quadrant analysis in identifying moderators.

In general, moderated regression has not been successful in improving predictions. Differences in results among moderator techniques are a consequence of comparing techniques that treat the same variable differently; that is, either as a quantitative or qualitative variable. Problems arise in that some techniques are appropriate for discrete qualitative variables, and other techniques are appropriate for continuous quantitative variables. A moderated regression equation (Saunders, 1956) describes the multivariate regression surface with a model that is nonlinear in the predictor (including moderator) variables. The techniques leading to the differential validity or differential predictability consider each subdivision of the original population to have a different linear relationship between predictor and criterion. Thus, the effect of any nonlinearity in the relationship between the moderator and the predictor-criterion combination may be ignored. As a result, close correspondence between results of quantitative and qualitative techniques would be dependent chiefly on the skillfulness in choosing the composition of the subgroups. In addition, statistical and methodological problems limit comparisons.

Problems with Moderator Techniques

Statistical and Methodological Problems

Several general statistical and methodological problems, particularly in the differential validity and predictability techniques, are apparent in the moderator approaches. The first problem, which is pertinent to all moderator approaches, is that of cross-validation.

Some investigators have reported results that were not cross-validated (e.g., Dawis et al., 1967; Murray[14]). Banas and Moore[15] presented a study that claimed to demonstrate the need for cross-validation. A sample of foremen was trichotomized into subgroups based on job satisfaction scores and, also, trichotomized randomly. (Whether the random assignment of foremen actually produced a different subgroup composition than that based on job satisfaction was not reported.) Three cases, each with different combinations of predictors, were examined to determine whether job satisfaction moderated prediction of performance. Results indicated that the patterns of validity coefficients were similar for both types of trichotomization. Consequently, Banas and Moore concluded that cross-validation is necessary in moderator research.

This writer agrees with the argument for cross-validating research. The results and conclusions of the Banas and Moore (see

---

[13] Zedeck, S., Cranny, C. J., Vale, C. A., & Smith, P. C. A comparison of "joint moderators" in three prediction techniques. Unpublished manuscript, Bowling Green University, Ohio, 1969.

[14] Murray, S. L. Moderators of training satisfaction and training performance relationships. Paper presented at the meeting of the Rocky Mountain Psychological Association, Albuquerque, New Mexico, May 1969.

[15] Banas, P. A., & Moore, L. Moderating effect of job satisfaction and evaluating moderator effectiveness. Paper presented at the meeting of the Midwestern Psychological Association, Chicago, May 1968.

Footnote 15) study, however, are confounded by a second problem, which is especially relevant to differential validity and differential predictability studies. This problem concerns the test for demonstrating differential validities among qualitative subgroups or predictability subgroups. Investigators have used three methods for demonstrating differential validities: (*a*) Significantly different coefficients *between subgroups;* that is r_{XY} for Subgroup 1 is significantly different from r_{XY} for Subgroup 2; (*b*) patterns of validity coefficients reveal some coefficients to be significantly different from *zero;* that is, r_{XY} for Subgroups 1 and 2 is significantly different from zero, but r_{XY} for Subgroup 3 is not significantly different from zero; and (*c*) the validity coefficient of a *subgroup* is different *from* the validity coefficient of the *total* group; that is, r_{XY} for Subgroup 1 is greater than r_{XY} for the total group, whereas r_{XY} for Subgroup 2 is equal to or less than r_{XY} for the total group.

The first way of demonstrating differential validities is more powerful than the second way. The significance from zero of subgroup coefficients is not fundamental to the question of differential validity. Consider the case where the coefficient for one subgroup is .40, and for the other subgroup the coefficient is .42, both significantly different from zero. Or consider the case when .42 is significantly different from zero, but .40 is not. In either event, can we conclude that the predictor is useful for one group and not for another?; that is, can we conclude that the moderator variable has yielded differential validities? The Banas and Moore (see Footnote 15) study, in addition to others (e.g., Dawis et al., 1967; Kirkpatrick et al., 1968) are examples of such statistical reasoning. On the other hand, if two validity coefficients are significantly different from each other, then the predictor can be more appropriate for one subgroup than the other.

As for the third way, this author does not know of a statistic that tests for the difference of a subgroup from the coefficient for the total group. Pervin (1967), Prediger (1966), and Grooms and Endler (1960), for example, formed conclusions based on such "analyses," but did not report the statistical test that was applied. Logically, this type of test would be most desirable to determine whether moderators are adding to prediction. Yet, if we compare the two validity coefficients, the best we can do is to "eyeball" the differences.

Perhaps different statistics should be used in assessing differential coefficients. McNemar (1969) has demonstrated that different validity coefficients do not necessarily imply that criterion scores for one group can be predicted more accurately than for the other subgroup. The subgroup with the highest validity coefficient does not necessarily have the smallest standard error of estimate. Consequently, Einhorn and Bass (1969) argued that the differences in errors of estimate for different subgroups, in addition to validity coefficients, should be examined. Another way to assess differential validities is to examine expectancy tables for the different subgroups (Guion, 1965, 1966). Different expectancy tables require different evaluations and interpretations of the data.

A third general problem concerns the number of different subgroups that should be formed. The possibility for forming an endless number of subgroups exists, particularly when examining biographical information blanks or demographic variables as moderators. Infinite subgrouping is related to the problems of sample size (subgroup sample sizes become too small on which to perform significance tests) and stability of the differences. Solutions are to consider several moderators simultaneously as the basis for subgroups or to use a single multiple-moderated-regression equation, which maintains the total sample size in the analyses.

A fourth problem pertains to the validity coefficient of the original total sample. Improving upon a validity coefficient that is already relatively high is both difficult and somewhat impractical. While the basic purpose of any moderator variable approach is to improve validity in situations in which predictors are poor, whether a gain is worth the effort involved in applying a technique, is a matter to be decided by the individual making the selection decisions or by a decision process that includes the notion of utility.

A fifth general problem is that the effect of the potential moderator variable as an independent predictor often is not reported (e.g., Brown, see Footnote 9; Murray, 1969, see Footnote 14). Applying the potential moderator variable as an independent predictor in an ordinary multiple regression equation can produce equally good results and thus would eliminate the endless subgrouping and subgroup sample size problems. Also, there would be less claims for the ubiquitous moderator in the research literature. Hence, the correlations, slopes, and standard error of estimates should be examined for the following three regression equations:

$$y = a + bx, \qquad [1]$$

$$y = a + b_1x + b_2z, \qquad [2]$$

where z is the potential moderator but is treated as an independent predictor, and

$$y = a + b_1x + b_2z + b_3xz \qquad [3]$$

(moderated regression equation). If Equations 2 and 3 are significantly different from Equation 1, but not from each other, then the variable is an independent predictor and not a moderator variable.

Operation of Moderator Variables

The literature concerning moderator variables has focused primarily on demonstrating the existence of moderators rather than explaining how they operate. A common conception is that moderators operate by sorting heterogeneous aggregates of individuals into homogeneous subgroups (Berdie, 1961; Johnson, see Footnote 2; Saunders, 1956). The subgroups are intended to be homogeneous with respect to error and psychological structure, with the magnitudes and patterns of reliability and validity coefficients varying from group to group (Baker, 1967). Banas (see Footnote 3) stated, however, that subgrouping is not simply homogeneous grouping but is grouping restricted usually to the extreme scores on the moderator variable.

According to Ghiselli (1963), evidence indicates that moderators differentiate those individuals in a group for whom error of measurement or prediction is small from a group for whom it is large. This explanation presumes that individuals can be divided into clear and distinct classes. Ghiselli claimed that, in actual practice, moderators distribute individuals along a continuum. Individuals are sorted into separate classes, and a group consists of individuals who fall at the same point on the continuum.

Another possible explanation of moderator effects is that the common elements, which account for the correlation between two variables, differ from individual to individual rather than from group to group (Ghiselli, 1963). Rather than sorting into classes or groups, each falling at a single point on the continuum, there is sorting into class intervals. Such a notion presumes that error of measurement varies from small for some individuals to large for others. Consequently, error scores would carry less weight in determining observable scores for some individuals than for others. However, a necessary condition is that individual differences in error scores possess reliability over parallel tests.[16]

Likewise, error of prediction is smaller and test validity higher for individuals at one extreme on the moderator continuum, and, at the other extreme, error of prediction is larger and test validity lower (Ghiselli, 1963). Consequently, the weight a test carries in prediction varies from individual to individual. With respect to validity, the function of the moderator is to predict for a given individual the weight a test carries in determining criterion performance. However, nothing in the concept indicates the relationship of the individual's weights to the criterion and/or test scores.

The above conceptions are appropriate depending on the given situation. Little attention, however, has been given to the mathematical operation of moderators. In one study, Conger (1969) developed and then unsuccessfully tested the implications of the mathematical bases of Ghiselli's (1956) predictability variable. The study (Conger, 1969) suggested that the predictability variable reflects deviations from the norm on some trait rather than deviations from the trait itself (e.g., presence or absence of the trait).

---

[16] Interestingly, though the notion of error of measurement is used to explain the operation of moderator variables, error of measurement with respect to the moderator itself is ignored. Rarely is the reliability of the moderator reported, and it would not be surprising if large error of measurement was responsible for moderator results not holding up in cross-validation.

The crucial determinant of the operation of moderators, as in the assessment of efficiency of moderators, depends on the type of variable, discrete or continuous, examined. Conceptualizations of moderator variables involve the distinction between continuous and discontinuous qualitative variables. The argument made by Ghiselli in his 1968 study and in an unpublished study[17] is that the moderator does not function to separate a heterogeneous population into a more homogeneous one. A *true* moderator is not a discontinuous qualitative variable that differentiates subgroups of individuals who are qualitatively different, but is a continuous quantitative variable, and individuals distribute all along its continuum. For qualitatively different groups the pattern of correlations can be different, and classic psychometric theory still is appropriate, for there is no moderator effect. If groups differ along a continuum, however, then a new psychometric theory is needed to elicit the moderator effect. In essence, there is a continuous (linear or curvilinear) relationship between scores on the moderator and some predictor-criterion score combination. This conceptualization is similar to Saunders' (1956) position that a moderator variable is a generalized case of analysis of covariance, with the basic parameter a score on a continuous variable rather than group membership.

In situations in which discontinuous qualitative variables have been used as moderators (e.g., subgroup analysis), a heterogeneous population is divided into more homogeneous subgroups. A complex aggregate of individuals is simplified, and there is no interaction between the moderator and predictor-criterion score combination.

Marks (1964) also has related off-quadrant techniques to analysis of variance. A significant interaction between predictor and criterion indicates the possibility that a moderator is operating and that the off-quadrants should be examined.

Off-quadrant techniques are explained in another way. Lykken and Rose (1963) suggested that the predictability of the criterion (Y) from the predictor (X) varies as a function of a moderator (Z), which may be uncorrelated with either Y or X. The function relating Y to X differs among individuals in the sample. One equation, for instance, $Y = aX$, is appropriate for the group close to the perfect regression line, whereas another equation, for instance, $Y = b/X$, is appropriate for those further from the regression line (the off-quadrant cases). The predictability of Y from X varies as a function of Z. Z is a function of the likelihood that an individual belongs to the $Y = aX$ group; that is, Z is a discriminant function.

Hobert and Dunnette (1967) applied the Lykken and Rose (1963) explanation to the quadrant analysis approach. Hobert and Dunnette (1967) claimed that equation $Y = aX$ is appropriate for the low hits and high hits, whereas the underpredicted and overpredicted are described by $Y = b/X$. In this case, Z is correlated with Y. Also, two equations are necessary to describe the off-quadrant cases.

Zedeck (see Footnote 11) applied discriminant analysis in the multipredictable group validation model with little success. The discriminant variables (moderators) identified in the experimental groups were not effective in the cross-validation group.

Consequently, there are several alternative suggestions as to how moderators operate. Yet, moderators are as elusive as suppressor variables. In fact, we may consider a suppressor variable as one type of moderator, though the reverse is not necessarily true. A suppressor variable is correlated with the predictor and uncorrelated with the criterion, whereas a moderator is uncorrelated with both the predictor and criterion.

---

[17] Ghiselli, E. E. The prediction of human behavior. Paper presented at the meeting of the Canadian Psychological Association, Calgary, June 1968.

Identification of Moderator Variables

A review of the moderator variable literature indicates that a considerable number of scales and variables function as moderators. The problem of readily and systematically identifying a moderator variable, however, still remains unsolved. Banas (see Footnote 3) distinguished between the rational and empirical approach to the identification of moderators.

The rational approach to the identification of moderators involves intuition, hypothesis-forming, and/or accident. Saunders (1955) realized the problem of identifying moderators and suggested four procedures: (*a*) Start with a good hypothesis; (*b*) study a few cases intensively and seek out variables of which interpretation seems to depend on other variables; (*c*) look for subgroups within which correlations are significantly different; and (*d*) test the interaction variance of a pair of items against a criterion. Toops (1959) suggested that moderators are found by looking for "if . . . provided" statements. Frederiksen and Melville (1954) examined compulsiveness as a moderator, for it "appeared to be useful" as the basis for a hypothesis.

Conger (1969) suggested four approaches for looking for moderators: (*a*) Look for differences between current subjects and standardization groups; (*b*) look for differences in item characteristics and formats of the tests used; (*c*) look for subject characteristics of a stylistic nature; and (*d*) look for measures of degree of deviation from a norm. These methods of trying out specific hypotheses, however, are time consuming, inefficient, and costly because of the considerable number of possibilities that exist.

The techniques by Ghiselli (1956, 1960b), Berdie (1961), and Hobert and Dunnette (1967), for example, are considered empirical approaches. Banas (see Footnote 3) mentioned several empirical approaches: (*a*) correlating variables with absolute differences, (*b*) correlating items with the difference between actual and predicted criterion, (*c*) assessing intraindividual variability, or (*d*) assessing statistical infrequency of response and/or response inconsistency. Though the techniques approach objectivity, one variable at a time is investigated, and, therefore, the procedure is not efficient.

Consequently, statistics or procedures that systematically or mechanically identify moderators are needed. The task is difficult, since not enough evidence has accumulated to make it possible to state any general principles about the nature of items or traits that function as moderators.

There have been few empirical attempts to solve the identification problem. French[18] attempted to develop a program for the selection of moderator variables from a large number of variables. The program was designed to identify moderators from 42 potential moderator variables without taking the time to compute actual correlations. Indexes, based on the joint distribution, in 3 × 3 contingency tables, of predictor and criterion scores that had been trichotomized, were used in place of correlation coefficients. The technique was unsuccessful because of distortion of the indexes.

Rock, Barone, and Linn (1967) developed a computer program for a moderated stepwise prediction system. Potential moderators are examined simultaneously or in a stepwise manner. Klein, Rock, and Evans (1968) had limited success with the program in one study. They favored the stepwise system over factorial designs because the latter require an impractically large number of groups, which are difficult to obtain in educational and industrial settings. In addition, factorial designs lose information by not treating the moderator as a continuous variable.

Zedeck (see Footnote 10) attempted to use discriminant analysis and analysis of covariance to identify moderators from among 17 potential variables in the multipredictable group validation model. The assumption was that the significant high weights of the discriminant function were operating as moderators. Cross-validation of the moderators identified by the approach proved unsuccessful.

The need for systematic identification of moderators is still evident. Abrahams (1969), in a summary statement, claimed that investigators were looking elsewhere to enhance prediction since moderators were being found only as often as chance permitted. Though the author sees no evidence for reduction in the number of moderator studies, the utility of the techniques will diminish unless a systematic technique is developed to identify moderators from a large mass of data. It is possible, as Ghiselli (1963) cautioned, that the time and effort required to develop and find moderators might be more productively spent improving reliability and validity of the sort that follows from classic theory.

---

[18] French, J. W. *A machine search for moderator variables in massive data.* (Office of Naval Research Report, Contract Nonr 2214(00)-NR 151-174) Princeton, N. J.: Educational Testing Service, 1961.

UTILITY OF MODERATOR VARIABLES

Consideration of the problems with moderator techniques raises the question of the utility of the approaches. The usefulness of the concept is evident in situations in which a validity coefficient is low, but, where at least with certain individuals, reasonably accurate prediction of criterion performance may be made from scores on the test. Selection experts can make use of different prediction equations and different patterns of validities for different links in the prediction models suggested by Guetzkow and Forehand (1961) and Dunnette (1963). Though the number of studies concerned with moderating effects is great, the results and problems limit their utility.

Though predictions are improved for one subgroup, there remains usually another subgroup for whom predictive efficiency is diminished. The implications are to eliminate them or to find new predictors for them. Since variables such as race, color, sex, and religion often are found to be moderators, the former implication is illegal and unethical. Therefore, greater attention must be given to the search for valid predictors for all groups in the population.

Weitz's (1966) "sequential moderator" notion might enhance the utility of the moderator concept and alleviate the problem of rejecting those individuals for whom the predictor appears to be inappropriate. The typical validation approach calls for validating a test and examining the influence of moderators at a particular point in time. It is recognized that the pattern of individual differences associated with performance changes with practice over time (Fleishman & Hempel, 1954). Thus, the time at which criterion (e.g., performance) measures are taken should be varied and, consequently, the predictor (e.g., measures of motivation or personality) validated at the appropriate time as dictated by a moderator (e.g., aptitude). We might establish from sequential moderator research when predictors are most valid for different subgroups.

Dodd, Wollowick, and McNamara (1969) tested the sequential moderator concept. Results indicated that whereas a personality measure was related to salary in early years for a group with high training grades, personality was unrelated for a group with low training grades for the same period. However, in later years, personality predicted equally well for both high and low training-grade subgroups. Dodd et al. concluded that training grade or task difficulty is a sequential moderator. We might consider also that years and training grades were "joint moderators" (as discussed below). However, the notion suggests that we do not have to eliminate subgroups for whom it is initially difficult to predict. In the long run, the success of predicted and unpredicted individuals may be equally predictable. Furthermore, having long-range estimates of success can prevent individuals with potential from leaving or being eliminated.

Another way to increase the utility of the moderator concept is to consider the operation of joint moderators. Joint moderators are defined as two or more variables that interact to influence the relationship between predictor and criterion (Zedeck et al., see Footnote 12). Fredericksen and Melville (1954) indicated that considerable improvement in prediction was obtained for a subgroup considered noncompulsive based on two different indexes. Stricker (1966) formed four subgroups from the combination of the two dichotomized compulsivity measures. In this case, the joint use of the two variables did not enhance the moderator effect. This result may be attributable to the fact that both moderators were measures of the same construct.

Klein et al. (1968) used the procedure suggested by Rock et al. (1967) to examine the effectiveness of joint moderators in the prediction of law school grades. Five items (age, major field of study, degree of preparation by college for law school, father's occupation, and time at which decision was made to attend law school) were used to form multiple moderators. The grouping procedure yielded four homogeneous groups. The group with the highest coefficient was significantly different from two of the remaining three groups.

Zedeck et al. (see Footnote 12) examined the operation of two joint moderators, study habits and anxiety, in the moderator models of subgroup analysis, prediction of predictability (Ghiselli, 1956), and moderated regression (Saunders, 1956). None of the techniques provided improved prediction with the use of joint moderators. The lack of results in this study and the limited success in others may be attributable to the types of variables used to form the joint groups. In addition, in the latter study, the correlation between the two variables may have contributed to the ineffectiveness of the joint use. Finally, the problem of subgroup sample size is more evident in joint moderators than when moderators are examined individually. In general, the results of joint moderator research further demonstrate the need for systematic identification of moderators.

Zedeck's (see Footnote 11) research on multipredictable group validation with discriminant analysis essentially considers the simultaneous operation of moderators (discriminant scores). However, as already noted, the results were unclear. Perhaps using factor analysis to diminish the number of potential variables and, then, factor scores to form predicted, overpredicted, and underpredicted groups would enhance prediction. It is possible that the factor structure varies considerably from group to group.

Tesser et al. (1967) computed factor scores derived from life history items for their subjects, but then placed them into homogeneous groups based on an index of profile similarity rather than based on degree of predictability. The results were inconclusive because of nonsignificance in cross-validating; however, lack of significance in cross-validation might have been because of extremely small sample sizes.

CONCLUSION

The utility of the moderator variable research is limited by statistical problems, by the limited understanding of the statistical operation of moderators, and by lack of a rapid systematic approach to the identification of moderators. With respect to the statistical problems, comparisons between the standard error of estimate, slope, and intercept for the subgroups should be made. Different statistics are needed since a correlation coefficient can indicate different patterns of relationships, which are a result of nonnormal distribution, skewness, or restrictions in range. Perhaps, to alleviate methodological problems resulting from subgrouping, moderated regression equations for the total group can be extended to include squared and other higher-order terms, thus allowing detectability of other forms of nonlinearity.

Finally, there is the semantics problem. The term moderator is used ambiguously in the literature. As demonstrated, the term may refer to control or population variables, or interactions, and/or differences in predictability, depending on the procedure used. Though the terms do not have entirely independent meanings, we should keep the mathematical and methodological meanings distinct.

REFERENCES

Abelson, R. P. Sex differences in predictability of college grades. *Educational and Psychological Measurement,* 1952, **12**, 638–644.

Abrahams, N. M. Off-quadrant comment. *Journal of Applied Psychology,* 1969, **53**, 66–68.

Berdie, R. F. Intra-individual variability and predictability. *Educational and Psychological Measurement,* 1961, **21**, 663–676.

Berdie, R. F. Consistency and generalizability of intraindividual variability. *Journal of Applied Psychology,* 1969, **53**, 35–41. (a)

Berdie, R. F. Intra-individual temporal variability and predictability. *Educational and Psychological Measurement,* 1969, **29**, 235–257. (b)

Carroll, S. J., Jr., & Tosi, H. L. Relationship of goal-setting characteristics as moderated by personal and situational factors to the success of the "management by objectives" approach. *Proceedings of the 77th Annual Convention of the American Psychological Association,* 1969, **4**, 599–600. (Summary)

Cleary, T. A. An individual differences model for multiple regression. *Psychometrika,* 1966, **31**, 215–224.

Conger, A. J. An analysis of Ghiselli's moderator variable. *Psychological Reports,* 1969, **25**, 519–527.

Dawis, R. V., Weiss, D. J., Lofquist, L. H., & Betz, E. Satisfaction as a moderator in the prediction of satisfactoriness. *Proceedings of the 75th Annual Convention of the American Psychological Association,* 1967, **2**, 269–270. (Summary)

Dodd, W. E., Wollowick, H. B., & McNamara, W. J. *Task difficulty as a moderator of long-range prediction.* (Experimental Publication System Issue No. 2; Ms No. 051A) Washington, D. C.: APA, EPS, 1969.

Doll, R. E., & Gunderson, E. K. E. Occupational group as a moderator of the job satisfaction-job performance relationship. *Journal of Applied Psychology,* 1969, **53**, 359–361.

Dunnette, M. D. A modified model for test validation and selection research. *Journal of Applied Psychology,* 1963, **47**, 317–323.

Dunnette, M. D., & Kirchner, W. K. Psychological test differences between industrial salesmen and retail salesmen. *Journal of Applied Psychology,* 1960, **44**, 121–125.

Einhorn, H. J., & Bass, A. B. *Discrimination in employment testing.* (Experimental Publication System Issue No. 2; Ms. No. 048A) Washington, D. C.: APA, EPS, 1969.

Fleishman, E. A., & Hempel, W. E., Jr. Changes in factor structure of a complex psychomotor test as a function of practice. *Psychometrika,* 1954, **19**, 239–252.

Frederiksen, N., & Gilbert, A. C. F. Replication of a study of differential predictability. *Educational and Psychological Measurement,* 1960, **20**, 759–767.

Frederiksen, N., & Melville, S. D. Differential predictability in the use of test scores. *Educational and Psychological Measurement,* 1954, **14**, 647–656.

French, J. W. The relationship of problem-solving styles to the factor composition of tests. *Educational and Psychological Measurement,* 1965, **25**, 9–28.

Gaylord, R. H., & Carroll, J. B. A general approach to the problem of the population control variable. *American Psychologist,* 1948, **3**, 310. (Abstract)

Ghiselli, E. E. Differentiation of individuals in terms of their predictability. *Journal of Applied Psychology,* 1956, **40**, 374–377.

Ghiselli, E. E. Differentiation of tests in terms of the accuracy with which they predict for a given individual. *Educational and Psychological Measurement,* 1960, **20**, 675–684. (a)

Ghiselli, E. E. The prediction of predictability. *Educational and Psychological Measurement,* 1960, **20**, 3–8. (b)

Ghiselli, E. E. Moderating effects and differential reliability and validity. *Journal of Applied Psychology,* 1963, **47**, 81–86.

Ghiselli, E. E. *The validity of occupational aptitude tests.* New York: Wiley, 1966.

Ghiselli, E. E. Interaction of traits and motivational factors in the determination of the success of managers. *Journal of Applied Psychology,* 1968, **52**, 480–483.

Ghiselli, E. E., & Sanders, E. P. Moderating heteroscedasticity. *Educational and Psychological Measurement,* 1967, **27**, 581–590.

Goodstein, L. D., & Heilbrun, A. B., Jr. Prediction of college achievement from the Edwards Personal Preference Schedule at three levels of intellectual ability. *Journal of Applied Psychology,* 1962, **46**, 317–320.

Grooms, R. R., & Endler, N. S. The effect of anxiety on academic achievement. *Journal of Educational Psychology,* 1960, **51**, 299–304.

Guetzkow, H., & Forehand, G. A. A research strategy for partial knowledge useful in the selection of executives. In R. Tagiuri (Ed.), *Research needs in executive selection.* Cambridge: Harvard University Press, 1961.

Guion, R. M. *Personnel testing.* New York: McGraw-Hill, 1965.

Guion, R. M. Employment tests and discriminatory hiring. *Industrial Relations,* 1966, **5**, 20–37.

Hobert, R., & Dunnette, M. D. Development of moderator variables to enhance the prediction of managerial effectiveness. *Journal of Applied Psychology,* 1967, **51**, 50–64.

Hoyt, D. P., & Norman, W. T. Adjustment and academic predictability. *Journal of Counseling Psychology,* 1954, **1**, 96–99.

Hull, C. L. *Aptitude testing.* Yonkers, New York: World Book, 1928.

Kipnis, D. A noncognitive correlate of performance among lower aptitude men. *Journal of Applied Psychology,* 1962, **46**, 76–80.

Kirkpatrick, J. J., Ewen, R. B., Barrett, R. S., & Katzell, R. A. *Testing and fair employment.* New York: New York University Press, 1968.

Klein, S. P., Rock, D. A., & Evans, F. R. The use of multiple moderators in academic prediction. *Journal of Educational Measurement,* 1968, **5**, 151–159.

Kogan, N., & Wallach, M. A. *Risk taking: A study in cognition and personality.* New York: Holt, Rinehart & Winston, 1964.

Korman, A. K. Self-esteem as a moderator of the relationship between self-perceived abilities and vocational choice. *Journal of Applied Psychology,* 1967, **51**, 65–67.

Korman, A. K. Self-esteem as a moderator in vocational choice: Replication and extensions. *Journal of Applied Psychology,* 1969, **53**, 188–192.

Lawler, E. E. III. Ability as a moderator of the relationship between job attitudes and job performance. *Personnel Psychology,* 1966, **19**, 153–164.

LYKKEN, D. T., & ROSE, R. Psychological prediction from actuarial tables. *Journal of Clinical Psychology*, 1963, **19**, 139–151.

MARKS, M. R. How to build better theories, tests, and therapies: The off-quadrant approach. *American Psychologist*, 1964, **19**, 793–798.

MCNEMAR, Q. Moderation of a moderator technique. *Journal of Applied Psychology*, 1969, **53**, 69–72.

PERVIN, L. A. Aptitude, anxiety and academic performance: A moderator variable analysis. *Psychological Reports*, 1967, **20**, 215–221.

PETERSON, F. E. Identification of sub-groups for test validation research. *Journal of Industrial Psychology*, 1964, **2**, 98–101.

PREDIGER, D. J. Application of moderated scoring keys to prediction of academic success. *American Educational Research Journal*, 1966, **3**, 105–111.

PREDIGER, D. J. Development of moderated scoring keys for psychological inventories. *Educational and Psychological Measurement*, 1969, **29**, 813–824.

RICHARDSON, H. Utility of a new method for predicting college grades. *Journal of General Psychology*, 1965, **72**, 159–164.

ROCK, D. A., BARONE, J. L., & LINN, R. L. A fortran computer program for a moderated stepwise prediction system. *Educational and Psychological Measurement*, 1967, **27**, 709–713.

SAUNDERS, D. R. The "moderator variable" as a useful tool in prediction. *Proceedings of the 1954 invitational conference on testing problems*. Princeton, N. J.: Educational Testing Service, 1955.

SAUNDERS, D. R. Moderator variables in prediction. *Educational and Psychological Measurement*, 1956, **16**, 209–222.

STEINEMANN, J. H. Use of a logically related predictor in determining intragroup differential predictability. *Journal of Applied Psychology*, 1964, **48**, 336–338.

STRICKER, L. J. Compulsivity as a moderator variable: A replication and extension. *Journal of Applied Psychology*, 1966, **50**, 331–335.

TESSER, A., STARRY, A. R., & CHANEY, F. B. Toward better prediction: A subgrouping approach. *Proceedings of the 75th Annual Convention of the American Psychological Association*, 1967, **2**, 261–262. (Summary)

TOOPS, H. A. A research utopia in industrial psychology. *Personnel Psychology*, 1959, **12**, 189–225.

WAGNER, M. E., & STRABEL, E. Homogeneous grouping as a means of improving the prediction of academic performance. *Journal of Applied Psychology*, 1935, **19**, 426–446.

WEITZ, J. Criteria and transfer of training. *Psychological Reports*, 1966, **19**, 195–210.

(Received March 31, 1970)

Manuscripts Accepted for Publication in the *Psychological Bulletin*

Task Structure and the Experimental Modification of Verbal Behavior: G. Alan Marlatt*: Department of Psychology, University of Wisconsin, Charter at Johnson, Madison, Wisconsin 53706.

Gene Action and Behavior of C3H and DBA Mice: A Reply to Wilcock and Thiessen: Daniel Bovet*, Filomena Bovet-Nitti, and Alberto Oliverio: Laboratorio di Psicobiologia e Psicofarmacologia del C. N. R., Via Reno 1, 00198, Rome.

Type IV Errors and Interactions: Joel R. Levin* and Leonard Marascuilo: Department of Educational Psychology, University of Wisconsin, 1404 Regent Street, Madison, Wisconsin 53706.

Nonexistent Individual Differences in Reminiscence: Edward N. Peters*: University of Rochester School of Medicine, Well Baby Research Unit, 1545 Mount Hope Avenue, Rochester, New York 14620.

Comment on Joe's "Comment on Overall and Spiegel's 'Least Squares Analysis of Experimental Data' ": I. Leon Smith*: 206 Teachers College, University of Cincinnati, Cincinnati, Ohio 45221.

Comment on Interpretation of Latitude of Rejection as an "Artifact": Carolyn W. Sherif*: Pennsylvania State University, 417 Psychology Building, University Park, Pennsylvania 16802.

Rejoinder to Sherif: O. W. Markley*: Western Behavioral Sciences Institute, 1150 Silverado, La Jolla, California 94025.

Olfactory Control of Behavior in Rodents: E. Fred Schultz and Jack T. Tapp*: Department of Psychology, Vanderbilt University, Nashville, Tennessee 37203.

Autokinetic Illusion: A Systematic Review of Theories, Measures, and Independent Variables: John Levy*: 12 Birriga Road, Bellevue Hill, Sydney, N. S. W., Australia, 2023.

---

* Asterisk indicates author for whom the address is supplied.

VOL. 76, NO. 5 NOVEMBER 1971

Psychological Bulletin

DESENSITIZATION:

SOCIAL AND COGNITIVE FACTORS UNDERLYING THE EFFECTIVENESS OF WOLPE'S PROCEDURE

WALLACE WILKINS[1]

University of Washington

Recent evidence suggests that Wolpe's theoretical formulations are inadequate in accounting for the effectiveness of the therapeutic approach of systematic desensitization. Of the elements traditionally included in the desensitization procedure, neither training in muscle relaxation nor the construction of a graded hierarchy of fear-relevant scenes nor the concomitance of instructed imagination to muscle relaxation are necessary conditions for treatment success; instructed imagination by itself is the only necessary element of the desensitization procedure. Further, the effectiveness of the procedure does not appear to be due to the traditionally stated mutual antagonism between muscle relaxation and anxiety, but rather due to social variables involved in the patient-therapist relationship and cognitive variables involving expectancy of therapeutic gain, information feedback of success, training in the control of attention, and vicarious learning of the contingencies of behavior through instructed imagination.

The use of systematic desensitization as a psychotherapeutic treatment method has grown rapidily since its introduction by Wolpe (1958) and currently enjoys widespread popularity among many mental health disciplines. Lazarus (1963) has documented its effectiveness in treating a wide variety of disorders, and Rachman (1967) thoroughly reviewed the research on systematic desensitization available at that time, described the procedure and the appealing theoretical rationale for its effectiveness. A number of recent investigations have been reported that, while not disclaiming the effectiveness of the procedure, do call into serious question the theoretical underpinnings from which Wolpe (1958) developed this technique.

The most popular form of desensitization employs muscle relaxation as an inhibitor of fear or anxiety, and, according to Wolpe, the stages of the procedure are (*a*) training in deep muscle relaxation, (*b*) arranging fear-provoking scenes into a graded hierarchy, and (*c*) the instructed imagination of the hierarchy scenes concomitant to a state of muscle relaxation. Recent studies, taken together, suggest that of the three above conditions, none are necessary for successful desensitization. Indeed, the only necessary condition appears to be imagination of fear-relevant scenes not necessarily arranged into a hierarchy and not necessarily concomitant to muscle relaxation. Thus, Wilson and Smith (1968) reported success in severely neurotic subjects where treatment consisted of relaxation contiguous to the imagination of free-association scenes that were not organized into a hierarchy. Certainly, individually tailored hierarchies are unnecessary, as Emery and Krumboltz (1967) found no difference in anxiety reduction whether they employed individualized hierarchies or a standard hierarchy developed by the investigators prior to

---

[1] I wish to thank Hans O. Doerr for his significant contributions in the development of this article. Requests for reprints should be sent to the author who is now at Department of Psychology, University of Miami, Coral Gables, Florida 33124.

© 1971 by the American Psychological Association, Inc.

interviewing subjects. Leitenberg, Agras, Barlow, and Oliveau (1969) also demonstrated fear reduction using a standard nonindividualized hierarchy for all subjects. Rachman (1968) presented arguments that training in muscle relaxation appears unnecessary, as successful fear reduction has been demonstrated with experimenters who are fairly inexperienced in the technique of muscle relaxation and with subjects who receive only perfunctory relaxation training. Also, Garfield, Darwin, Singer, and McBrearty (1967), Cooke (1966), and Ritter (1968) reported successes with *in vivo* desensitization where subjects are physically active during treatment sessions. Lazarus (1965) reported successful cases where subjects are instructed to increase muscle activity concomitant to imagining fear-provoking scenes. Davison (1966) demonstrated that muscle relaxation and anxiety are not mutually antagonistic, since anxiety can be experienced while a person is in a state of chemically produced muscle relaxation. Finally, Wolpin and Raines (1966) demonstrated successful fear reduction in a group of subjects who imagined hierarchy scenes with no instructions about relaxation, another group who imagined hierarchy scenes while intentionally tensing their muscles, and a third group who received no instructions regarding relaxation and imagined only scenes at the top of the hierarchy with no graded exposure to less fear-provoking scenes. Wolpin (personal communication, 1966) also reported successful case studies in which imagination of fear-provoking scenes was used without relaxation training and without arranging them into a graded hierarchy.

These studies strongly suggest that Wolpe's theoretical formulations are inadequate in identifying the critical variables involved in treatment outcome. Of the elements included in Wolpe's procedure, instructed visualization of hierarchy scenes appears to be the only one necessary for treatment success. The reports by Wolpin and Raines (1966) and Wolpin (personal communication, 1966) indicate further that instructed visualization may also be a sufficient condition for therapeutic gain. The remainder of this article focuses on factors operating in the desensitization situation which research evidence indicates are highly influential in affecting the outcome of treatment.

Expectancy of Therapeutic Gain and the Therapist as a Reinforcer of Nonfear Behavior

A study by Wagner and Cauthen (1968) is suggestive of the effect the therapist has as a social reinforcing agent. "Treatment" did not focus on fear; instead, subjects received a liberal amount of compliments for their appearance and attitudes. When given verbal reinforcement for approaching a snake, they showed reduction in avoidance behavior.

Leitenberg et al. (1969) found significant differences in fear reduction between control subjects and subjects who received relaxation contiguous to imagination of hierarchy scenes, instructions forming an expectation of therapeutic gain, and praise for small improvements. However, no differences were found between no-treatment controls and subjects who received relaxation contiguous to imagination of hierarchy scenes, instructions designed to eliminate expectation of therapeutic gain, and no praise for improvements. The authors argued that since the combination of relaxation plus imagination of hierarchy scenes did not produce significant differences from the control group, the major therapeutic benefit from desensitization was due to the instructed expectation of success or verbal reinforcement for small successes.

These two effects, expectation of therapeutic gain and reinforcement for small improvements, were separated by Oliveau, Agras, Leitenberg, Moore, and Wright (1969) in a 2 × 2 factorial design such that one group received both, a second group received instructions forming an expectation of therapeutic gain, a third group received praise for small improvements, and a fourth group received neither. All groups also received relaxation contiguous to imagination of hierarchy scenes. Although all groups showed improvement, an analysis of variance indicated that only the factor involving expectation of therapeutic gain had a significant effect; praise for improvement and the interaction factors were not significant.

From these studies it appears that the skill of the therapist to develop an expectancy of therapeutic gain and to act as a social reinforcing agent for improvement contributes in a large part to therapeutic gain. These variables in the therapeutic relationship appear to be quite powerful in that they can have an immediate and long-term effect following very minimal contact with the therapist (see Kahn & Baker, 1968; Lang, 1969; Migler, 1968; Migler & Wolpe, 1967; and Rardin, 1969 for procedures involving desensitization with minimal therapist contact).

The Role of Information Feedback

Providing a subject with information about his performance has therapeutic effects, as shown by Leitenberg, Agras, Thompson, and Wright (1968). Two phobics were merely asked to remain in the presence of anxiety-producing stimuli until they felt anxious. The authors found that when the subjects were given precise information as to how long they had remained in the presence of the stimuli, the length of exposure time increased and performance variability decreased as compared to trials when this information was not given. Praise added to information had little effect.

Snake-fearing subjects in studies conducted by Valins and Ray (1967) heard heartbeat-sounding noises increase in tempo during presentations of shock stimuli and remain at base level during presentations of snake stimuli. In reality the tempo was controlled by the experimenter; however, subjects who were told that the noises were from their own heart and who were, in effect, receiving information that they were not aroused during presentations of snake stimuli showed greater reduction in avoidance behavior than did subjects who were told that the sounds were just extraneous noises.

In the desensitization procedure, there are several possible sources of information feedback that may contribute to change in fear reduction. Of these, the subject's perception of his own behavior, progression up the hierarchy, and physiological reactions are discussed.

1. *Subject's perception of his own behavior as information feedback.* Bem (1967) has marshaled evidence that self-reports of attitudes and feelings are, in part, a function of the self-observation of overt behavior. A self-report of an internal event is viewed as an inference from observing one's own overt behavior in the presence of accompanying situational variables—subjects are using the same information (overt behavior) that an outside observer would use in making inferences about a subject's attitudes or emotions. For illustration, Bandler, Madaras, and Bem (1968) applied shock under conditions where subjects were either instructed to escape or endure the shock. Even though shock intensity level was the same, subjects rated escaped shock as significantly more painful than endured shock. The authors offered the interpretation that the reports of pain were controlled by subjects' perception of their own overt behavior—if they observed themselves escaping, pain was reported to be greater than if they observed themselves enduring.

This interpretation is applicable to the desensitization situation: If a subject endures the imagination of a scene, he should conclude that he is more comfortable and should report less anxiety than if he observes himself signaling to terminate the scene imagination. A variety of factors may be influencing signaling behavior (e.g., reinforcement for not signaling, patient's susceptibility to therapist's persuasion, demand characteristics of the situation). The point is that the behavior of signaling to terminate scene imagination is present or absent. Its presence or absence serves as information that can influence a subject's self-reports of fear and changes in avoidance behavior. The information function of overt behavior operates even when the meaning of signaling is reversed as in the successful procedure presented by Cautella (1966) where subjects are instructed not to signal the presence of anxiety, but only when the anxiety is no longer experienced to the imagined scene.

2. *Progression up the hierarchy as information feedback.* An additional source of information, when it is included in treatment, is the perception of progression up the hierarchy. Lang and Lazovik (1963), Lang, Lazovik, and Reynolds (1965), Davison (1968), and Cooke (1968) have found that the number of hierarchy scenes passed is highly related to fear reduction. Subjects who pass all

or nearly all of the hierarchy scenes show greater fear reduction than subjects who do not come as near to completing the hierarchy in the same number of sessions. The latter group is typically no different after treatment than subjects who receive no treatment.

As presented earlier, the use of a graded hierarchy is not necessary in producing successful desensitization; however, when included in the treatment procedure, a subject's progression up the hierarchy does provide him with information of improvement. Self-perceived progression through the hierarchy may function as data like other self-perceived behaviors, which may influence reports of fear and changes in avoidance behavior.

3. *Perceived physiological reactions as information feedback.* Rachman (1968), Zeisset (1968), Leitenberg et al. (1969), Lomont and Edwards (1967), and Doerr (personal communication, 1969) reported studies in which physiological activity of subjects receiving treatment has been recorded. No significant differences on the physiological level have been found between treatment and control subjects or between pre- and postmeasures in subjects receiving successful treatment, even though significant differences are observed in behavioral and self-report measures of fear. Following treatment, the desensitized group sometimes shows less reduction than untreated subjects in physiological activity recorded in the presence of the previously feared stimuli. This casts serious doubts on the notion that physiological reactions serve an information function influencing reduction in fear and avoidance behavior.

An avenue for accounting for the findings—that through desensitization, changes in emotions and emotional behavior consistently occur without concomitant changes on the physiological level—is offered by the empirical and theoretical developments advanced by Schachter and his associates (Bramel, Bell, & Margulis, 1965; Mandler, 1962; Nisbett & Schachter, 1966; Schachter, 1964; Schachter & Singer, 1962; Schachter & Wheeler, 1962; Valins, 1966; Valins & Ray, 1967). Their work demonstrates that emotional states are jointly determined by physiological and cognitive-perceptual factors: Given the presence of physiological arousal, an individual will label or describe this state in terms of the cognitions available to him. These investigators present evidence that the same physiological state may be experienced, labeled, and described as anger, joy, fear, anxiety, love, or a variety of other emotions, depending on cognitive-perceptual factors involving the situation in which the arousal is experienced.

If cognitive factors are major determinants of the affect experienced in a given situation, it becomes a distinct possibility that the emotion (and subsequently, approach or avoidance behavior) can be altered by providing a subject with experiences through which the cognitive label applied to the feeling state is altered. A subject could show the same pattern of physiological activity in the presence of a previously feared stimulus, but through cognitions acquired in treatment he could be labeling his arousal in a different way and therefore be experiencing a different emotion. New cognitions available to a subject are, for example, that he has been able to imagine scenes involving the stimulus without signaling the presence of fear or anxiety. Fear then is an inappropriate label for stimulus-produced autonomic activity. From personal observations of approach behavior testing, emotional expressions of fear prior to treatment are changed following treatment to expressions of excitement and glee at being able to actually handle the previously feared stimulus.

The Role of Instructed Imagination

Evidence was presented at the beginning of this study that instructed imagination of feared stimuli is the only necessary element of Wolpe's procedure. There appear to be two aspects of this process through which subjects learn ways to respond to the feared stimuli other than escaping or avoiding it:

1. *Controlled attention shifts.* The benefit of procedures involving changes in muscle tension concomitant to scene imagination may not be due to reciprocal inhibition of anxiety, but due to training subjects in a technique that allows a termination of attention directed toward the threatening situation. In the conventional desensitization, the therapist's sequence of instruction during imagination of scenes is relax, imagine, relax, stop imagining,

relax. There is an instructed shift of attention from muscles, to threatening scene, back to muscles, allowing repeated practice in voluntary termination of attention directed toward the fear-producing stimuli. For reasons of experimental design, termination of scene imagination was not under the control of subjects in one of Davison's (1968) groups. This group showed little fear reduction, and the author suggested that self-control of exposure to feared stimuli may be a critical variable in successful desensitization. Parallel to this, Mowrer and Viek (1948) demonstrated that animals who could control the termination of a noxious stimulus developed fewer signs of anxiety than did subjects receiving the same exposure but for whom termination of the stimulus was not under their control.

Other fear-reduction techniques that do not use relaxation do contain the characteristic of repeated practice in shifting attention from threatening imaginal stimuli. Lazarus' (1965) technique of directed muscle activity concomitant to imagination of fear-provoking scenes causes a disruption of attention directed toward the stimuli. The emotive imagery technique introduced by Lazarus and Abramovitz (1962) involves the instructed practice of controlling attention between threatening and pleasant imaginal stimuli. Wolpe (1960) reported a decrease in frequency of an obsessive thought by applying, contiguous to the imagination of the scene, a mild shock which caused a disruption in the imagined scene.

2. *Exposure to contingencies of nonavoidant behavior in the fear-provoking situation.* In avoiding an aversive situation, a subject also misses exposure to information about consequences of nonavoidant behavior in that situation. Experiences through which a subject learns nonavoidant techniques appropriate to the anxiety-provoking situation should help alleviate fear and avoidance behavior in that situation.

Direct instruction of behavior appropriate to a threatening situation is, as demonstrated by Zeisset (1968), one manner in which contingencies can be learned. He found that relaxation training and instructions in how to invoke relaxation in an anxiety-provoking situation were as effective in reducing anxiety as the full desensitization procedure.

In vivo experiences with the aversive stimuli offer another form of nonavoidant contingency learning. Dogs in a study by Seligman, Maier, and Geer (1968) learned coping behavior in an extremely aversive situation after the experimenters physically forced the subjects to make the coping responses. Cooke (1966) and Garfield et al. (1967) demonstrated desensitization in humans when they were encouraged to respond nonavoidantly to anxiety-producing stimuli presented to them *in vivo*.

Vicarious learning by observing a model engaged in nonavoidant behavior with feared stimuli is another contingency learning experience. Bandura, Grusec, and Menlove (1967) and Ritter (1968) demonstrated fear reduction after subjects watched a peer approach the feared stimulus.

A similar modeling process can be theorized to occur during the conventional desensitization process when subjects are instructed to imagine hierarchy scenes. As such, a subject may be, in effect, "modeling" after his own imagined behavior. Although there is no direct empirical support for this notion, it is interesting to note that the scenes typically imagined during desensitization are not of the feared stimulus per se, but rather they are scenes in which the subject imagined himself behaving in some fashion in relation to the feared stimulus (e.g., stepping on a live snake or touching a snake). The subject imaginally observing himself successively approach the feared stimulus may very well be a process parallel to the subject observing a live model engaged in these successive approach behaviors.

SUMMARY

A decade ago both the therapeutic procedure of systematic desensitization and its theoretical underpinnings were introduced. The procedure has been established as a very effective technique for fear and anxiety reduction; however, in light of investigations reported predominantly in the last two years, the original theoretical formulations appear to be inadequate in providing an understanding of the nature of the learning involved in

desensitization. Strong trends in the research evidence indicate that neither relaxation training nor hierarchy construction are necessary for successful therapy outcome. It appears that the only necessary element of the desensitization procedure is the cognitive element of instructed imagination of fear-relevant scenes.

It appears that the cognitive and social aspects of the therapeutic situation, which are common both to desensitization as well as to more traditional forms of psychotherapy, are critical variables for successful therapeutic outcome. These variables include expectancy of therapeutic gain, the social reinforcing qualities of the therapist, information feedback of approximations toward successful fear reduction, training in the control of attention, and the vicarious learning (via instructed imagination) of contingencies of nonavoidance behavior in the fear situation.

REFERENCES

Bandler, R. J., Madaras, G. R., & Bem, D. J. Self-observation as a source of pain. *Journal of Personality and Social Psychology*, 1968, **9**, 205–209.

Bandura, A., Grusec, J. E., & Menlove, F. L. Vicarious extinction of avoidance behavior. *Journal of Personality and Social Psychology*, 1967, **5**, 16–23.

Bem, D. Self perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, 1967, **74**, 183–200.

Bramel, D., Bell, J. E., & Margulis, S. T. Attributing danger as a means of explaining one's fear. *Journal of Experimental Social Psychology*, 1965, **1**, 267–281.

Cautella, J. R. A behavior therapy approach to pervasive anxiety. *Behaviour Research & Therapy*, 1966, **4**, 99–109.

Cooke, G. The efficacy of two desensitization procedures: An analogue study. *Behaviour Research & Therapy*, 1966, **4**, 17–24.

Cooke, G. Evaluation of the efficacy of the components of reciprocal inhibition psychotherapy. *Journal of Abnormal Psychology*, 1968, **73**, 464–467.

Davison, G. C. Anxiety under total curarization: Implications for the role of muscular relaxation in the desensitization of neurotic fears. *Journal of Nervous and Mental Disease*, 1966, **143**, 443–448.

Davison, G. C. Systematic desensitization as a counterconditioning process. *Journal of Abnormal Psychology*, 1968, **73**, 91–99.

Emery, J. R., & Krumboltz, J. O. Standard versus individualized hierarchies in desensitization to reduce test anxiety. *Journal of Counseling Psychology*, 1967, **14**, 204–209.

Garfield, Z. H., Darwin, P. L., Singer, B. A., & McBrearty, J. F. Effect of "in vivo" training on experimental desensitization of a phobia. *Psychological Reports*, 1967, **20**, 515–519.

Kahn, M., & Baker, B. Desensitization with minimal therapist contact. *Journal of Abnormal Psychology*, 1968, **73**, 198–200.

Lang, P. J. The on-line computer in behavior therapy research. *American Psychologist*, 1969, **24**, 236–239.

Lang, P. J., & Lazovik, A. D. Experimental desensitization of a phobia. *Journal of Abnormal and Social Psychology*, 1963, **66**, 519–525.

Lang, P. J., Lazovik, A. D., & Reynolds, D. J. Desensitization, suggestibility, and pseudotherapy. *Journal of Abnormal Psychology*, 1965, **70**, 395–402.

Lazarus, A. A. The results of behavior therapy in 126 cases of severe neurosis. *Behaviour Research & Therapy*, 1963, **1**, 67–79.

Lazarus, A. A. A preliminary report on the use of directed muscular activity in counter-conditioning. *Behaviour Research & Therapy*, 1965, **2**, 301–303.

Lazarus, A. A., & Abramovitz, A. The use of "emotive imagery" in the treatment of children's phobias. *Journal of Mental Science*, 1962, **108**, 191–195.

Leitenberg, H., Agras, W. S., Barlow, D. H., & Oliveau, D. C. Contribution of selective positive reinforcement and therapeutic instructions to systematic desensitization therapy. *Journal of Abnormal Psychology*, 1969, **74**, 113–118.

Leitenberg, H., Agras, W. S., Thompson, L. E., & Wright, D. E. Feedback in behavior modification: An experimental analysis in two phobic cases. *Journal of Applied Behavior Analysis*, 1968, **1**, 131–137.

Lomont, J. F., & Edwards, J. E. The role of relaxation in systematic desensitization. *Behaviour Research & Therapy*, 1967, **5**, 11–25.

Mandler, G. Emotion. In R. W. Brown, E. Galanter, E. H. Hess, & G. Mandler (Eds.), *New directions in psychology*, Vol. 1. New York: Holt, Rinehart & Winston, 1962.

Migler, B. A supplementary note on automated self-desensitization. *Behaviour Research & Therapy*, 1968, **6**, 243.

Migler, B., & Wolpe, J. Automated self-desensitization: A case report. *Behaviour Research & Therapy*, 1967, **5**, 133–135.

Mowrer, O. H., & Viek, P. An experimental analogue of fear from a sense of helplessness. *Journal of Abnormal and Social Psychology*, 1948, **43**, 193–200.

Nisbett, R. E., & Schachter, S. The cognitive manipulation of pain. *Journal of Experimental Social Psychology*, 1966, **2**, 227–236.

Oliveau, D. C., Agras, W. S., Leitenberg, H., Moore, R. C., & Wright, D. E. Systematic desensitization, therapeutically oriented instructions and selective positive reinforcement. *Behaviour Research & Therapy*, 1969, **7**, 27–33.

Rachman, S. Systematic desensitization. *Psychological Bulletin*, 1967, **67**, 93–103.

Rachman, S. The role of muscular relaxation in desensitization therapy. *Behavior Research & Therapy*, 1968, **6**, 159–166.

Rardin, M. W. Treatment of a phobia by partial self-desensitization: A case study. *Journal of Consulting and Clinical Psychology*, 1969, **33**, 125–126.

Ritter, B. The group desensitization of children's snake phobias using vicarious and contact desensitization procedures. *Behaviour Research & Therapy*, 1968, **6**, 1–6.

Schachter, S. The interaction of cognitive and physiological determinants of emotional state. In L. Berkowitz (Ed.), *Advances in experimental social psychology*, Vol. 1. New York: Academic Press, 1964.

Schachter, S., & Singer, J. Cognitive, social and physiological determinants of emotional state. *Psychological Review*, 1962, **69**, 379–399.

Schachter, S., & Wheeler, L. Epinephrine, chlorpromazine and amusement. *Journal of Abnormal and Social Psychology*, 1962, **65**, 121–128.

Seligman, M. E. P., Maier, S. F., & Geer, J. H. Alleviation of learned helplessness in the dog. *Journal of Abnormal Psychology*, 1968, **73**, 256–262.

Valins, S. Cognitive effects of false heartrate feedback. *Journal of Personality and Social Psychology*, 1966, **4**, 400–408.

Valins, S., & Ray, A. A. Effects of cognitive desensitization on avoidance behavior. *Journal of Personality and Social Psychology*, 1967, **7**, 345–350.

Wagner, M. K., & Cauthen, N. R. A comparison of reciprocal inhibition and operant conditioning in the systematic desensitization of a fear of snakes. *Behaviour Research & Therapy*, 1968, **6**, 225–227.

Wilson, A., & Smith, F. J. Counterconditioning therapy using free association: A pilot study. *Journal of Abnormal Psychology*, 1968, **73**, 474–478.

Wolpe, J. *Psychotherapy by reciprocal inhibition.* Stanford: Stanford University Press, 1958.

Wolpe, J. Reciprocal inhibition as the main basis of psychotherapeutic effects. In H. J. Eysenck (Ed.), *Behaviour therapy and the neuroses.* New York: Pergamon Press, 1960.

Wolpin, M., & Raines, J. Visual imagery, expected roles and extinction as possible factors in reducing fear and avoidance behavior. *Behaviour Research & Therapy*, 1966, **4**, 25–37.

Zeisset, R. M. Desensitization and relaxation in the modification of psychiatric patients' interview behavior. *Journal of Abnormal Psychology*, 1968, **73**, 18–24.

(Received April 17, 1970)

Psychological Bulletin
1971, Vol. 76, No. 5, 318–327

IS "BEHAVIOR THERAPY" BEHAVIORISTIC?

(AN ANALYSIS OF WOLPE'S PSYCHOTHERAPEUTIC METHODS)

EDWIN A. LOCKE[1]

University of Maryland

Wolpe's psychotherapeutic methods are analyzed from the point of view of whether or not they are, in fact, based on behaviorist assumptions. It is shown that Wolpe's procedures (*a*) require the patient to introspect and to reveal the content and intensity of his negative emotional states; (*b*) presuppose the patient's ability to think, reason, remember, judge, discriminate, and imagine (all of which are operations of consciousness); and (*c*) assume that the patient can purposefully control (certain of) the actions of his own mind as well as his overt behavior. It is concluded that Wolpe's procedures contradict every major premise of behaviorism. Some explanations for this breach between theory and practice are offered.

The term "behavior therapy" was coined by Skinner and Lindsley[2] to describe an approach to psychotherapy based on behavioristic learning theory. The term has since been widely promulgated by Eysenck (1960, 1964) and is now used to describe a variety of techniques, all of which are (allegedly) based on behavioristic principles.

Broadly speaking, "behaviorism is the doctrine that the behavior of man and animal can be fully understood without the use of explanatory concepts referring to states or actions of consciousness, namely by studying only observable behavior [Locke, 1969, pp. 1000–1001]." The basic premises of behaviorism are (*a*) *determinism:* the doctrine that all of man's actions, thoughts, beliefs, etc., are ultimately determined by forces outside his control (typically, by environmental forces); (*b*) *epiphenomenalism:* the doctrine that conscious states (e.g., ideas), if they exist at all, are merely by-products of physical events in the body and/or in the external world—that they are end points in a causal chain and have no effect on either the individual's subsequent actions or his subsequent ideas; and (*c*) the *rejection of introspection* as a scientific method. A behaviorist investigator may ask individuals to give "verbal reports," but such reports may not be used to make inferences regarding the subject's mental states or processes.[3]

The validity of the doctrine of behaviorism must be judged on the basis of whether or not the principles it espouses enable one to understand man and to account for (and help to modify) his actions. Thus, for example, if a self-proclaimed behavior therapist asserts that he can cure neurosis, it is relevant to ask (*a*) whether he can, in fact, achieve such cures; and (*b*) whether he achieves them by applying the principles he claims to be applying and using and the methods he claims to be using (i.e., behavioristic principles and methods).

The present study deals specifically with the second question. The purpose of this article is to analyze the psychotherapeutic techniques of one proponent of behavior theory—Joseph Wolpe—from the point of view of whether or not his methods are in fact behavioristic. Wolpe's system was chosen for analysis for two reasons: (*a*) His theoretical assumptions and therapeutic procedures are described thoroughly in the literature; and (*b*) his procedures appear to be among the most widely used by those "behavior" therapists who treat neurosis.

---

[1] The author acknowledges the helpful comments of William Claiborn of the University of Maryland, and Arthur Mode of Stamford, Connecticut, although the views expressed are the sole responsibility of the author. Requests for reprints should be sent to the author, Department of Psychology, University of Maryland, College Park, Maryland 20742.

[2] Skinner, B. F., & Lindsley, O. R. Studies in behavior therapy. Status Reports II and III. (Contract N5 ori-7662) Washington, D. C.: Office of Naval Research, 1954.

[3] Skinner (1964) has recently argued that one should not reject a priori the study of private events (images, sensations, etc.) simply because they are private. This is a radical departure from strict behaviorism; however, in actual practice Skinner does not use introspection and considers mental events to be trivial and irrelevant in explaining human behavior.

Breger and McGaugh (1965) have previously criticized behavior therapy with respect to the validity of its assumptions, the meaningfulness of its concepts and definitions, its effectiveness in curing neurosis, and the relationship between theory and practice. The latter issue, however, was not treated in detail. While Breger and McGaugh's study was extensive in its coverage, in that it focused on many aspects of therapy and on three different behavior models, the present study is an intensive analysis of one aspect of a single approach to therapy.

The question of the efficacy of Wolpe's procedures is too broad to be dealt with here, but it will be granted, for the purposes of this study, that Wolpe is able to achieve certain beneficial results with certain patients, for instance, phobia remission. The question is: How does he do it?

To establish the fact that Wolpe claims to be a behaviorist, let us examine his formal theoretical system.

WOLPE'S FORMAL THEORETICAL SYSTEM

Wolpe's main theoretical concepts, for instance, "conditioned reflex," "conditioned inhibition," "drive reduction," "reinforcement," are all drawn from well-known behaviorists such as Pavlov, Watson, Hull, and Skinner. Wolpe views behavior as determined by external stimulation in conjunction with events in the nervous system. He argues that "the patient has . . . no choice in becoming what he is [Wolpe & Lazarus, 1968, p. 16]." Wolpe further asserts:

> every content of consciousness depends upon specific, though no doubt complex, neural activity. [Other researchers have observed] . . . that stimulation of points on the superior or lateral surfaces of the temporal lobe aroused imagery of past experiences . . . Any contents of the subject's consciousness would be *in parallel and not in series* with their neural correlates, and would in essence constitute the unique reaction of a specially placed observer (the subject) to these neural events [Wolpe, 1958, p. 16].[4]

It is clear that Wolpe views conscious experience as the by-product of brain activity, that is, as an epiphenomenon.

Consistent with his behaviorist premises, Wolpe does not attempt to define neurosis as such but rather defines *neurotic behavior*, which is

> *any persistent habit of unadaptive behavior acquired by learning in a physiologically normal organism.* Anxiety is usually the central constituent of this behavior, being invariably present in the causal situations [Wolpe, 1958, p. 32].

His formal definition of anxiety is also purely behavioral:

> By anxiety is meant the autonomic response pattern or patterns that are characteristically part of the organism's response to noxious stimulation [Wolpe, 1958, p. 34].

> The usual manifestations [of anxiety] are tachycardia, raised blood and pulse pressures, hyperpnea, pilo-erection, mydriasis, palmar hyperidrosis, and dryness of the mouth [Wolpe, 1958, p. 35].

According to Wolpe "anxiety responses" are acquired through a process of conditioning, for instance, the subject is hurt by some physically noxious stimulus, and his subsequent "fear response" to that situation then generalizes automatically to other (similar) situations. Such responses can be removed by "conditioning" some other (positive) response to the same situation. All of this is asserted to be totally outside the control of the subject himself (cf. Wolpe & Lazarus, 1968, p. 16), although no proof for this claim is given.

---

[4] The finding that direct stimulation of a person's brain produces memory images does not prove that all acts and contents of consciousness are produced by antecedent brain activity over which the individual has no control. For it has also been found that when an individual sets himself a purpose to think of different types of events, one can record systematic changes in the electrical activity of his brain (Brown, 1970a, 1970b). A behaviorist who took the first finding as evidence that ideas were the end point of a causal chain beginning with brain activity would, to be consistent, have to conclude from the second finding that brain activity was the end point of a causal chain beginning with conscious awareness! Since these two positions are mutually contradictory, there is obviously some flaw in the deductive process. A more plausible position on the mind-brain relationship is that while consciousness presupposes a brain, the two are not identical, that is, the former is not reducible to the latter. For an excellent refutation of the doctrine of reductionism, see Blanshard, 1964, pp. 311–340.) Furthermore, in order to explain such activities as purposive thought and deductive reasoning, it is necessary to grant that conscious ideas *can* direct both subsequent brain activity and subsequent mental activity. (For a recent view along these lines, see Sperry, 1969.)

Wolpe's formal position on introspection is somewhat unclear but essentially behavioristic by default. While he suggests that one may use introspection to detect the autonomic effects of noxious stimulation (Wolpe, 1958, p. 34), nothing is said about its role in identifying conscious states or processes.

Having established that Wolpe's formal theoretical system is consistently behavioristic, let us now examine the actual procedures that he uses to treat patients.

WOLPE'S PSYCHOTHERAPEUTIC METHODS

This section is divided into two major parts, corresponding to the two aspects of psychotherapy: diagnosis and treatment.

Diagnosis

Judging from Wolpe's formal definition of neurosis, one would expect him to attempt to measure the patient's neurotic "behavioral responses" to various situations in order to discover his problem(s). Surprisingly, however, we find that the data used for diagnosis are derived from quite a different source. The patient is told to introspect and to report his characteristic negative feeling states (i.e., emotional reactions). For example, the patient is typically given the Willoughby questionnaire (Wolpe, 1958, p. 108), which asks the patient to indicate the frequency with which he experiences "fright," or "worry"; whether he is easily "hurt" or "discouraged"; if he feels "inferior" or "lonely"; and whether he has trouble "making up his mind." The patient may also be asked to complete a "fear inventory," which asks him to indicate what types of "things and experiences . . . cause fear or other unpleasant feelings [Wolpe & Lazarus, 1968, pp. 172–176]."

The information gained from these questions is supplemented with life history information, assigned homework, and interview data, all of which are aimed at revealing the *content* of the patient's negative feelings and emotions (Wolpe & Lazarus, 1968, pp. 24ff).

It should not be assumed from this that Wolpe merely uses the subjects' "verbal reports" as a shortcut method for measuring their physiological responses. Wolpe does *not* ask his patients (at this point) to tell him how tense their muscles feel, or whether their hands are sweating, or if their mouth feels dry; he asks them if they feel "anxious" or "disturbed" or "upset" or "worried." He is clearly interested in the conscious, experiential aspect of their emotions. This procedure is an implicit admission that emotions consist of something more than bodily responses (for an exceptionally thorough discussion of this issue, see Arnold, 1960). Even if the patient's physiological reactions (however measured) did accurately index the *type* of emotion felt (e.g., anxiety, anger, or guilt), they would not reveal the *content* (objects) of his emotions, that is, those situations and occurrences that arouse the emotions. Short of bringing every single object in existence into the therapist's office and measuring the patient's physiological response, the only way of determining a patient's characteristic emotional reactions to things is to ask him. To answer accurately, the patient must use introspection and memory; both actions require the operation of consciousness.

Usually the patient is also asked to indicate the relative *amount* of anxiety he experiences in different situations or in response to different objects or persons using a *subjective anxiety scale* (SUD scale; Wolpe & Lazarus, 1968, p. 73). Occasionally, the subject is asked to rate the emotional reactions that he would *expect* to have to situations that he has not actually experienced. To do this, he is told to *imagine* himself in the hypothetical situations and then to *imagine* (by inference) his reaction to them. Expectation and imagination are also actions of consciousness.

Nor can it be assumed that Wolpe is interested simply in the patient's "verbal report" as an end in itself. He mentions that in administering the Willoughby questionnaire the therapist may need to discuss the "implications" of each question with the patient before scoring his answer (Wolpe & Lazarus, 1968, p. 26). The reason for this is obviously that the therapist wants to know the *meaning* of the patient's response; he wants to make sure that the patient *understands* what is meant by the question before recording his answer. Wolpe reports a problem he had treating one patient who had not considered the anxiety she felt when imagining a certain situation to be "worthy of report [Wolpe & Lazarus, 1968, p. 71]." Therapy could not proceed successfully until the patient revealed her actual feelings. What the patient says is used as a *means* to an end, the end being the identification of the content and intensity of his negative emotional reactions.

It is clear from the foregoing that Wolpe's diagnostic methods all depend on the patient introspecting (and possibly imagining) and revealing the actual (or expected) content and intensity of his internal feeling states (emotions). These procedures could hardly be called behavioristic in view of the fact that they *all* pertain to the patient's conscious experiences.

Furthermore, the implicit premise behind Wolpe's procedure of asking the patient to reveal his emotional responses is that these emotional reactions *are* the patient's problem. It is assumed that they are the *cause* of his "unadaptive behavior," for instance, his avoidance, in real life, of situations that entail no objective threat. As is demonstrated below, therapy is aimed at modifying these feelings (and related beliefs) in order that the patient's actions may change.

Treatment

Preconditions of therapy. Wolpe asserts that successful treatment necessitates that the patient first be assured that his "unpleasant reactions" are reversible, that is, curable; "the therapist may find it desirable to offer advice, reassurance, support, encouragement or clarification [Wolpe & Lazarus, 1968, p. 27]." The therapist must also establish "a trustful relationship" and make the patient "feel fully accepted as a fellow human being [Wolpe & Lazarus, 1968, p. 28]." With respect to children, Wolpe's colleague, Lazarus, writes: "A warm rapport . . . and a close understanding of his wish-fulfillments and identifications are *essential* [italics added; Lazarus & Abramovitz, 1964, p. 461]." The implicit assumption behind such "supportive" actions is that they will influence the patient's conscious attitudes (toward the therapist and therapy, etc.) and that these attitudes will in turn affect the patient's actions, for instance, lead to his cooperation with the therapist.

Frequently, patients' "misconceptions" must be corrected before therapy can proceed successfully. Examples of mistaken beliefs that Wolpe has had to correct in his patients include the following: pine tree bark secretes poison (Wolpe, 1958, p. 200); masturbation is harmful (Wolpe & Lazarus, 1968, p. 131); direct genital contact is the only way to achieve orgasmic satisfaction (Wolpe & Lazarus, 1968, p. 105); and chest pains inevitably signify heart disease (Wolpe & Lazarus, 1968, p. 19). In order to treat one patient who suffered from intense irrational jealousy toward his wife, Wolpe had to "establish *a clear line of demarcation between situations that would give legitimate grounds for jealousy and those in relation to which jealousy would be abnormal* [Wolpe, 1958, p. 164]." To convince another patient that she was not mentally abnormal:

> The behavior therapist had to employ every resource of reason and demonstration and the full force of his own prestige to persuade her that she was in truth a normal person who had become disabled by neurotic conditionings . . . [Wolpe & Lazarus, 1968, pp. 131–132].

Other patients are convinced of the irrationality of their religious beliefs by doing assigned reading on the topic (Wolpe, 1969, p. 20).

The technique of persuasion is not confined simply to the patient's mistaken *beliefs;* it is also applied to his *moral values.* In order to get his patients to become more self-assertive, Wolpe frequently has to first convince them that abject self-sacrifice is self-destructive and that healthy self-assertion is morally proper. Wolpe writes:

> For patients with a strongly moralistic training logical arguments may be needed to convince them of the virtue of standing up for one's rights and the unavoidability (and justice) of hurting the feelings of those who flout those rights [Wolpe & Lazarus, 1968, p. 44].

One patient, for example, believed that she had no right to expect her aunt to stop taking advantage of her. Wolpe said to her: "But all you're expecting is justice. Is justice too much to expect?"[5]

---

[5] Wolpe, J. The Case of Mrs. Schmidt. Counselor Recordings, Nashville, Tennessee, 1965. (Transcribed tape recording)

Patients may also be provided with practical knowledge (such as how to take a job interview, or how to answer back to a husband or father, or how to deal with an irrational boss; see Wolpe & Lazarus, 1968, pp. 48–53), as well as practice in applying that knowledge. To accomplish the latter, the therapist and patient each take the role of one of the protagonists in the patient's real-life situation.

The procedures described above—which are precursors to successful therapy—consist in essence of correcting the patient's wrong ideas and providing him with new knowledge and/or values. This is known in the vernacular as *changing the patient's mind!* It is not well publicized as a behavioristic technique.

SPECIFIC THERAPEUTIC PROCEDURES

Only Wolpe's three major techniques are discussed here: self-assertion, systematic desensitization, and sexual arousal.

Self-Assertion

If the patient is diagnosed as being anxious and passive in the presence of other people, he is urged to become more self-assertive in his dealings with others:

> The therapist increases the motivation by pointing out the emptiness of the patient's fears, emphasizing how his fearful patterns of behavior have incapacitated him and placed him at the mercy of others, and informing him that, though expression of resentment may be difficult at first, it becomes progressively easier with practice. It usually does not take long for the patient to begin to perform the required behavior, although some need much initial exhortation and repeated promptings [Wolpe, 1958, p. 115].

Observe that the patient is motivated to initiate self-assertive behavior through reason and persuasion, not Pavlovian conditioning. Furthermore, the patient has to learn to distinguish between situations where assertion is appropriate and situations where it is not, a task that requires him to use judgment, discrimination, and thinking.

In his formal theory, Wolpe implies that the benefits of self-assertion are due solely to its inhibition of anxiety (Wolpe, 1958, p. 115), but Wolpe's own description of the effects of self-assertion is quite revealing on this point and worth quoting in full:

> Most assertive trainees seem to follow a similar pattern of evolution. First, there is an increased *awareness* of their non-assertiveness and its negative repercussions. This is followed by an *intellectual appreciation* of assertive behavior and its positive effects. Increasing distaste for their own ineffectuality and *resentment* towards the forces which seem to be maintaining or reinforcing the nonassertiveness soon lead to tentative, usually clumsy, attempts at self-assertive responses. If positive effects ensue, the probability of engaging in more assertive behavior increases. Occasionally, as *emotional satisfactions* intensify and spread, previously timid and dominated individuals tend to overassert themselves . . . [eventually] . . . The patient learns to be dominant without being dominating . . . Finally, as the patient becomes *aware* of his growing mastery of interpersonal situations, there develops a genuine and fitting *indifference* to minor slights, petty machinations, small irrationalities and other insignificant "pinpricks" of daily interaction. . . . an additional consequence of assertive training is a changed *self-concept.* More adequate behavior elicits positive feedback from other individuals, and this may modify existing *negative self-perceptions* in a way that facilitates the performance of the new behavior [italics added; Wolpe & Lazarus, 1968, pp. 47–48].

Note that the patient's "awareness," his "intellect," his "self-concept," and his emotions are all granted a role not only in the initiation but in the maintenance and subsequent modification of his self-assertive actions. Clearly, a great deal more is involved than mere inhibition of anxiety.

Systematic Desensitization

This procedure is used for patients whose fears are of noninterpersonal situations or for whom assertive training is not appropriate or adequate. The patient and therapist establish anxiety or fear hierarchies in the diagnosis sessions, and treatment begins by instructing the patient in muscle relaxation. (The question of whether it is only the patient's muscles that are relaxed by this procedure is dealt with below). Wolpe tells the patient to deliberately contract and then to deliberately relax specific muscle groups, until all parts of the body can be relaxed at will (Wolpe, 1969, p. 102). The implicit premise of this procedure is that the patient's conscious purposes affect the actions of his body (muscles)!

Once the patient has been taught to relax, he is instructed to *imagine* (picture in his mind) the object or situation in (one of) his fear hierarchies that produces the least amount of anxiety. This procedure implies that the patient's conscious purposes can influence the actions of his mind (his ideas)!

Both this and the premise underlying the training of muscle relaxation are not only at variance with but are *directly contradictory to* the doctrine of epiphenomenalism, which asserts that conscious ideas (e.g., purposes) are the end points of a causal chain and exert no independent influence on the individual's actions or ideas.[6]

As systematic desensitization proceeds, the patient informs the therapist of the amount of anxiety he is experiencing, a procedure that requires the patient to introspect. Again it should be stressed that Wolpe is interested not simply in the patient's verbal report but in the relation of that report to his real feelings (see Wolpe & Lazarus, 1968, p. 71).

With respect to one patient, Wolpe writes:

> It is worth remarking that even though the patient has a signal [lifting his finger] at his disposal with which to indicate [emotional] disturbance, the fact that he has not done so during a [an imagined] scene by no means proves that it has not disturbed him at all, for it is a rare patient who makes use of the signal if only mildly disturbed [Wolpe, 1964, pp. 28–29].

When treating another patient Wolpe told her repeatedly: "express your feelings [pp. 3–4, see Footnote 5]. Clearly, he viewed her feelings as something separate and distinguishable from her overt behavior, or else there would have been no sense in telling her to express them.

In sum, Wolpe is obviously interested in more than his patient's verbalizations. What the patient says (along with other evidence, e.g., his nonverbal actions, his past history, his facial expressions) is used as a basis for *inferring* his mental states. Wolpe uses what is directly observable as a basis for inferring what cannot be directly observed.

Sexual Arousal

Wolpe's method of treating patients with sexual problems (premature ejaculation, etc.) consists mainly of (*a*) *suggesting* new sexual procedures to the patient and (*b*) *persuading* the patient to abandon any irrational or impossible demands that he (or she) makes of himself. For example, a patient who is very anxious about his sexual performance might be told that he should go about sex with the attitude that "He has no duty at any stage to reach any criterion of performance [Wolpe, 1958, p. 131]." Pains are taken to make sure the partner is "understanding," "tolerant," and "affectionate."

The mechanism that Wolpe holds responsible for virtually all of the improvements made by his patients, regardless of the particular method used or the nature of the problem, is that of "reciprocal inhibition." Let us now examine this concept in some detail.

The Concept of Reciprocal Inhibition

As coined by Sherrington in 1947, the term reciprocal inhibition referred to a specific neurophysiological phenomenon, namely,

> the inhibition of one spinal reflex by another, such as occurs when stimulation of an ipsilateral afferent nerve causes relaxation of a vastocrureus muscle contracting to a contralateral stimulus [Wolpe, 1958, p. 29].

In other words, reciprocal inhibition occurs when one type of reflex inhibits the action of another type of reflex.

Wolpe, however, argues with respect to this term that

> Its use may be expanded to encompass all situations in which the elicitation of one response appears to bring about a decrement in the strength of evocation of a simultaneous response [Wolpe, 1958, p. 29].

According to this arbitrarily broadened definition, the following types of events would presumably be examples of "reciprocal inhibition": If a man is self-assertive in a given situation, he will not simultaneously be passive in that situation; if a man feels self-confident, he will not at the same time feel inferior; if a man gets pleasure from sex, he will not simultaneously fear it; if a man is relaxed in the company of others, he will not at the same time be tense; if a man turns left at a street corner, he will not turn right at the same time; if a man says "goodbye" to a friend, he will not simultaneously say "hello,"

---

[6] Opposition to the doctrine of epiphenomenalism (a doctrine that is self-contradictory, see Locke, 1966) does not require that one advocate the view that ideas can exist independently of a physical brain or that one believe in ghosts (see Footnote 4).

Lest this last example appear too farfetched, consider the following quote from Wolpe: "accompanying the articulation of any word there is ordinarily an automatic inhibition of all simultaneous tendencies to pronounce other words [Wolpe, 1960, p. 91]."

These examples assert, in essence, that one cannot do (or feel) two incompatible things at the same time. As Wolpe defines the term reciprocal inhibition, therefore, *it represents little more than a restatement* (in a narrower context) *of a fundamental metaphysical principle, which is identified in logic as the law of contradiction.* This law states that a thing cannot be both itself and not itself (A and non-A) at the same time and in the same respect. This principle was identified by Aristotle over 2,000 years ago.

Wolpe's assertion that reciprocal inhibition is the technique by which his patients lose their anxieties thus reduces to the claim that when a patient experiences one emotion other than (or incompatible with) anxiety in a given situation, he will not simultaneously experience anxiety in that same situation. While this is certainly true, it is not very scientifically enlightening. The relevant scientific issue is *how* to get patients to remain calm in situations that characteristically upset them.

Rachman (1968) has argued, on the basis of research evidence, that *muscle* relaxation is not a necessary element in successful desensitization. He argues that only *mental* relaxation may be necessary (and essential). On this point, one can observe in Wolpe's case accounts that he often instructs his patients to remain "calm." In "The Case of Mrs. Schmidt" (see Footnote 5), for example, after training the patient in muscle relaxation, Wolpe told her to raise her finger if anything affected her "state of calm." After each scene presentation, Wolpe asked Mrs. Schmidt if that scene had caused "worry" or "disturbance." He did *not* ask her if her muscles were tense. Calmness in this context clearly meant mental calmness.

More recently, Wilkins (1971) has argued on the basis of further evidence that relaxation may not be necessary at all for "desensitization" to be effective. Wilkins claims that it is sufficient to simply have the patient repeatedly imagine fear-producing scenes and to expect that the procedure will help him.

This author would suggest that Wolpe's actual psychotherapeutic procedures can be subsumed under five main categories: (*a*) convincing the patient that he can be helped by (behavior) therapy; (*b*) identifying the patient's irrational beliefs, values, and fears; (*c*) providing the patient with new knowledge and/or values; (*d*) helping and persuading the patient to act on this new knowledge; and (*e*) teaching the patient how to relax in the presence of formerly frightening situations (relaxation training, imagining situations, etc.). These procedures presuppose, explicitly or implicitly, that (*a*) the patient has the ability to introspect; (*b*) he has a mind capable of grasping concepts, understanding questions, and following logical arguments (i.e., that he can think); (*c*) he has the ability to purposefully control the actions of his own mind and his overt behavior.

Furthermore, it should not be assumed that the particular techniques Wolpe uses to modify a patient's irrational emotions are the only (or even the most desirable) methods that could be used to accomplish this goal. Blumenthal (1969), for example, has developed methods whereby the patient learns to "think his way out" of neurotic emotions, such as fear of disapproval. Wilkins (1971) suggests other possible methods.

CONCLUSION

The claim that Wolpe's psychotherapeutic methods are based on behavioristic principles cannot be supported in any important respect. Virtually everything that is done in the psychotherapy sessions contradicts one or more basic behavioristic premise. Nor can calling Wolpe's procedures behavioristic be defended on the grounds that his implicit conception of behaviorism is simply broader than that used by other behaviorists and is thus meant to *include* such activities as "feeling or emotional behavior," "thinking behavior," "understanding behavior," "imagining behavior," and "relaxing behavior." It must be stressed that behaviorism as a school and as a philosophical movement was a revolt *against the study of consciousness;* a strict behaviorist studies *only* observable behavior. To confess that one's implicit definition of behaviorism includes the study of consciousness is to confess that one is *not* a behaviorist. To broaden the term this much makes it both meaningless and unnecessary.[7]

There is one minor (and nonessential) respect in which Wolpe's method does resemble behaviorist procedures. He places considerable stress on *timing;* in systematic desensitization, for example, scene presentation must occur only when the patient is relaxed, that is, the disturbing scene is presented only (and immediately) after the patient is calm. In this sense the procedures resemble Pavlovian conditioning (where the unconditioned stimulus must be presented in close temporal proximity with the conditioned stimulus) and operant conditioning (where the "reinforcement" must follow immediately upon the desired response). In this respect, however, the methodology also resembles procedures used by (mental) associationists to study processes such as memory (presenting one nonsense syllable right after another).

The contradiction between behaviorist theory and the practice of behavior therapy by Wolpe cannot be resolved by asserting that words are "conditioned stimuli" that control behavior just like any other conditioned stimulus, for instance, as with a bell eliciting salivation in a dog. While word *sounds* (or any other sound) may under certain conditions come to serve as signals that elicit automatized responses, the effects of words qua *units of language* are far more complex. Words in the latter sense are symbols that stand for concepts; concepts represent an unlimited number of concrete objects, situations, or events (Rand, 1966). The effects of words or arguments on an individual depend upon such factors as how and if they are understood, how their meaning is appraised or evaluated, and the individual's decision to act. These mental operations are not mechanically conditioned from without; man has the capacity to direct (within limits) the operations of his own consciousness. He can initiate the acts of thinking and evaluation; he can set himself purposes and discover the means to achieve them. The understanding of enormously complex actions such as thinking and reasoning is not aided by arbitrarily attaching the label "conditioned stimuli" to the tools (words) man uses to accomplish these tasks.

Since the deviation of Wolpe's methods from genuine behavioristic techniques is so obvious and blatant, it is worth inquiring as to its causes.

The answer to the question of why Wolpe does not actually *use* behaviorist techniques seems obvious: if he were to use them, it would be impossible either to diagnose or to treat neurosis, that is, the procedures would not work. Such techniques might be applicable to the training of animals (and to an extent, to modifying trivial aspects of the behavior of autistic children and schizophrenics), but no therapist could adequately treat a neurotic patient without talking to him, a procedure that presupposes that the patient is conscious, that he understands words, that he can think and introspect.[8]

The issue of why, given that he does not actually use behavioristic methods, Wolpe still claims to do so, is more complex, but a plausible reason can be offered.

For decades psychologists have had an inferiority complex with respect to the physical sciences because an important aspect of their subject matter—man's mind—could not be studied by methods that other sciences had declared to be "scientific." Thus, psychologists have been accused (especially by each other) of being "mystics" rather than scientists, of studying "the ghost in the machine," of believing in "animism" and "primitive superstitions," and of "taking us back to the dark ages." In response to these criticisms many have attempted either (*a*) to reduce concepts referring to consciousness to concepts applicable to the physical sciences and thus study consciousness "in disguise," as it were; or (*b*) to dispense with the study of consciousness entirely and study only those aspects of behavior that *can* be studied by the methods of the physical sciences, that is, observable behavior. (For one example of the chaotic consequences of the first procedure, see Locke, 1969; for some documentation of the failure of the second, see Dulany, 1968; Spielberger & DeNike, 1966).

---

[7] Even if one were to grant (which this writer does not) that definitions are arbitrary, it is incumbent upon anyone who radically and fundamentally changes the definition of a concept he uses to publicly and openly announce this fact in order not to mislead his readers. To use the term behavior therapy, for example, without acknowledging the fact that one does not mean what everyone thinks he means by it, is to risk the danger that behaviorism (defined in the traditional way) will be given credit for something it does not deserve. This, of course, may well be the intent of those who misuse the term in this manner.

[8] One can argue that even with respect to animals and autistic children, one needs to posit some degree of conscious awareness on the part of the organism in order to *explain* why the so-called "reinforcer" reinforces. Descriptive behaviorism conveniently "avoids" the necessity for such explanations by defining reinforcement in terms of its effects and denying any interest in why reinforcers affect behavior (e.g., see Skinner, 1964, p. 104).

In short, psychologists have avoided studying or admitting they are studying consciousness due to fear of being labeled "unscientific" (see, on this point, the comments by Rychlak, 1969).

The irony of the situation is that the accusation is based on an "unscientific" conception of the nature of science. Science is the systematic study of the facts of reality. A rational scientist chooses methods of study that are appropriate to the nature of the entity or phenomenon being studied. He does not try to force such phenomena to fit into preconceived notions about methodology derived from the study of entirely different phenomena. A rational scientist, having observed (*a*) that consciousness is a crucial attribute of man and (*b*) that it cannot be studied or measured by the use of concepts and procedures employed in the physical sciences (e.g., consciousness cannot be weighed on a scale, or measured with a ruler, or chemically analyzed), would attempt to develop new methods that *were* appropriate for the study of this phenomenon. The failure to recognize this fact, in this writer's opinion, has not only produced enormous conceptual confusion in psychology, it has tremendously retarded the scientific study of consciousness.

In closing, there are two additional points that must be made with reference to behavior therapy: (*a*) The accusations made against Wolpe concerning the breach between theory and practice are not equally applicable to all advocates of behaviorism (e.g., Skinner [9]); and (*b*) the author's comments regarding the efficacy of Wolpe's techniques in helping neurotics to remove symptoms should not be taken as an implicit endorsement of his procedures. The issue of whether Wolpe's methods actually are applicable to all neurotics and the adequacy of these methods in helping the patient to attain full mental health are issues deserving of further comment but are not relevant to the purpose of the present study.

---

[9] Conventional behaviorists, however, do cash in on the fact that their subjects are conscious (see Footnote 8) but in a more subtle way than is the case with Wolpe.

REFERENCES

ARNOLD, M. *Emotion and personality*. Vol. 1. *Psychological aspects*. New York: Columbia University Press, 1960.

BLANSHARD, B. *The nature of thought*. Vol. 1. New York: Humanities Press, 1964.

BLUMENTHAL, A. The base of objectivist psychotherapy. *The Objectivist*, 1969, **8**, No. 6, 6–10; No. 7, 4–9.

BREGER, L., & MCGAUGH, J. L. Critique and reformulation of "learning-theory" approaches to psychotherapy and neurosis. *Psychological Bulletin*, 1965, **63**, 338–358.

BROWN, B. B. Awareness of EEG-subjective activity relationships detected within a closed feedback system. *Psychophysiology*, 1970, **7**, 451–484.

BROWN, B. B. Recognition of aspects of consciousness through association with EEG alpha activity represented by a light signal. *Psychophysiology*, 1970, **6**, 442–452. (b)

DULANY, D. E., JR. Awareness, rules and propositional control: A confrontation with S-R behavior theory. In D. Horton & T. Dixon (Eds.), *Verbal behavior and general behavior theory*. Englewood Cliffs, N. J.: Prentice-Hall, 1968.

EYSENCK, H. J. (Ed.) *Behavior therapy and the neuroses*. New York: Pergamon Press, 1960.

EYSENCK, H. J. (Ed.) *Experiments in behavior therapy*. New York: Macmillan, 1964.

LAZARUS, A., & ABRAMOVITZ, A. The use of "emotive imagery" in the treatment of children's phobias. In H. J. Eysenck (Ed.), *Experiments in behavior therapy*. New York: Macmillan, 1964.

LOCKE, E. A. The contradiction of epiphenomenalism. *British Journal of Psychology*, 1966, **57**, 203–204.

LOCKE, E. A. Purpose without consciousness: A contradiction. *Psychological Reports*, 1969, **25**, 991–1009.

RACHMAN, S. The role of muscular relaxation in desensitization therapy. *Behavior Research & Therapy*, 1968, **6**, 159–166.

RAND, A. *Introduction to objectivist epistemology*. New York: The Objectivist, Inc., 1967.

RYCHLAK, J. F. Lockean vs. Kantian theoretical models and the "cause" of therapeutic change. *Psychotherapy*, 1969, **6**, 214–222.

SKINNER, B. F. Behaviorism at fifty. In T. W. Wann (Ed.), *Behaviorism and phenomenology*. Chicago: University of Chicago Press, 1964.

SPERRY, R. W. A modified concept of consciousness. *Psychological Review*, 1969, **76**, 532–536.

SPIELBERGER, C. D., & DENIKE, L. D. Descriptive behaviorism versus cognitive theory in verbal operant conditioning. *Psychological Review*, 1966, **73**, 306–326.

WILKINS, W. Desensitization: Social and cognitive factors underlying the effectiveness of Wolpe's procedure. *Psychological Bulletin*, 1971, **76**, 311–317.

WOLPE, J. *Psychotherapy by reciprocal inhibition*. Stanford: Stanford University Press, 1958.

WOLPE, J. Reciprocal inhibition as the main basis of psychotherapeutic effects. In H. J. Eysenck (Ed.), *Behavior therapy and the neuroses*. New York: Pergamon Press, 1960.

WOLPE, J. The systematic desensitization treatment of neuroses. In H. J. Eysenck (Ed.), *Experiments in behavior therapy*. New York: Macmillan, 1964.

WOLPE, J. *The practice of behavior therapy*. New York: Pergamon Press, 1969.

WOLPE, J., & LAZARUS, A. *Behavior therapy techniques*. New York: Pergamon Press, 1968.

(Received May 19, 1970)

Psychological Bulletin
1971, Vol. 76, No. 5, 328–348

DEVELOPMENT OF THE HUMOR RESPONSE:

A REVIEW OF THE LITERATURE[1]

PAUL E. McGHEE[2]

University of Toronto

The aim of the present review is to summarize the empirical findings and theoretical views on children's humor and bring attention to methodological problems confronting a developmental approach to the study of humor. Arousal theory and classical psychoanalytic and gestalt theories are discussed, and a developmental theory of "cognitive" humor is advanced by the writer. While most of the studies reviewed deal with age differences in humor responsiveness or the relationship between intellectual development and children's humor, others deal with the humor of the exceptional or preschool child, sex differences, and the social nature of the humor situation. Specific attention is also given to unresolved theoretical and empirical issues relating to the development of the humor response.

In an early summary of research on humor, Piddington (1933) noted that philosophers and other writers have been speculating about the nature and importance of humor since Plato's time. Until the current century, however, the study of humor failed to advance from the armchair to the laboratory. As early as 1897, G. S. Hall (Hall & Allin, 1897) argued that theories of laughter or humor should go beyond mere speculation and be based on empirical data. Hall also felt that "there are few more promising fields of research." However, as may be seen in the more recent reviews of Flugel (1954) and Berlyne (1969) and the bibliography compiled by Treatwell (1967), the number of empirical studies of humor continues to be very modest. Furthermore, since most of these studies have focused on adult subjects, our knowledge of children's humor is especially meager. It is the purpose of this review to bring attention to our limited understanding of children's humor by summarizing both the theoretical positions that have been advanced and the empirical research that has been completed. Attention is also drawn to special methodological problems and important theoretical and empirical issues that confront a thorough investigation of children's humor.

THEORETICAL VIEWS

Only contemporary psychological theories of humor are summarized here, with special attention given to theoretical positions with specific developmental implications. A discussion of additional philosophical and psychological theories of humor may be found in Piddington (1933), Flugel (1954), and Berlyne (1969).

Psychoanalytic Theory

The most influential theory of humor to date continues to be that advanced by Freud in *Wit and Its Relation to the Unconscious* (1916). Freud distinguished between the comic, wit, and humor, emphasizing the cognitive, conative, and affective aspects of the ludicrous, respectively. In each case, the principle of "economy" of psychic energy plays the central role in producing the experience of satisfaction or pleasure. A "saving in thought" or representation characterizes the comic. Thus, comical situations involve a transition from a set to take something seriously to a perception that it is trivial or inconsequential. The consequent saving of energy is discharged as laughter. Freud considered wit to operate through plays on words or ideas, with satisfaction generated by a

---

[1] This manuscript was prepared during a postdoctoral year with Daniel E. Berlyne at the University of Toronto, supported by Department of Health, Education, and Welfare Grant HD 34940-01 from the National Institute of Child Health and Human Development.

[2] Requests for reprints should be sent to Paul E. McGhee, who is now at Department of Psychology, State University of New York at Albany, 1400 Washington Avenue, Albany, New York 12203.

"saving in inhibition." In "tendential" wit, normally inhibited sexual or aggressive ideas or urges find a temporary release through the socially sanctioned joke, and the heretofore pent-up instinctual energies find expression in the form of laughter. In "harmless" wit, pleasure is derived from a shedding of our normal efforts to maintain a moral, rational, and logical mode of cognitive functioning. While the overt expression of affect is likely to be more moderate here, a pleasurable delight accompanies such regressions to a more childlike indulgence in absurd or nonsensical thoughts and actions. Finally, the "saving in feeling" that characterizes humor is achieved by reducing the seriousness of a situation that would otherwise produce some strong negative emotion. By denying the threat, the individual feels better able to cope with the situation at hand. While the economy principle played an important role in Freud's thinking, Flugel (1954) noted that later psychoanalytic writers have not widely adopted it: "They have been inclined rather to emphasize the factor of discharge, the regression to infantile omnipotence, and the fact that in humor generally the discharge is under the control of the ego . . . [p. 721]."

While Freud did not devote special attention to humor in children, later psychoanalytic writers have extended his views along developmental dimensions. Wolfenstein (1954) argued that while the devices utilized vary with age, the basic motive in humor initiated by the child is the overcoming of distress and momentary release from frustration. Every child experiences frustration and anxiety from sources such as the making of impossible wishes and the overwhelming power and bigness of adults. Much of the children's joking is an attempt to free themselves from impossible longings and wishes by picturing their fulfillment as ridiculous. The use of fantasy is important throughout this process, since the child, in all of his play and joking, is pretending something is real while knowing it is not.

Wolfenstein asserted that as children develop they become increasingly aware of the restrictions placed on sexual and hostile impulses. Along with simply developing inner restraints, the child learns to use the joke as a means of expressing these impulses. Generally by the age of 6, the child sees the need for a joke facade; that is, a mechanism that creates an ambiguity between respectable and indecent aspects of the joke, thus allowing it to get past inner and outer censors and provide vicarious enjoyment.

Finally, Wolfenstein (1954) emphasized the role of mastery in children's humor by stressing that an important distinction must be drawn initially between whether the absurdity or nonsense of a joke is set forth by an adult or by the child himself: "While they can easily break away from the demands of reason on their own initiative and enjoy fantasy or nonsense, it is not always clear to them when others, especially their elders, make this shift [p. 196]." Only when the child has achieved a higher level of mastery and understanding of his environment does he begin to respond to the humor of adults' jokes. The earliness at which this occurs, according to Wolfenstein (1953), depends both on the child's intelligence level and his interest in jokes.

Kris (1938) also emphasized the importance of mastery, arguing that to experience the comic:

> A preliminary condition is complete control over that function in question. An absurd movement on the part of another person will seem funny to a child *only* when it has itself mastered the movement. At a later stage of development, it will laugh at a mistake in thinking, only when its own powers of thought are firmly established [p. 83].

Grotjahn (1957) argued that the child first discovers comic situations when he begins to master and enjoy body movements. When he begins to feel superior to other children in this respect, he is likely to see their mistakes or weaknesses as funny. Eastman's (1921) view that play contains the germ of the sense of humor is consistent with the position of Kris and Grotjahn, since the very young child's play consists primarily of motor behavior. While laughter appears to accompany a sheer joy of bodily movement in most infants, the mastery of motor skills through play probably plays an important role in the kind of superiority laughter described by Grotjahn and Kris. Levine (1967) also suggested that laughter and fun are inseparable from the young child's play. "It can safely be said that one cannot be truly playful without laughter, nor can one be humorous without being playful [p. 56]." Levine suggested that "both play and humor express a freedom from the realistic cares and problems of reality [p. 56]."

Helmers (1965) has advanced a position bearing some relation to the mastery theories described above. He maintained that the creation of nonsense and absurdity in children's jokes is really aimed at reconfirming the fact that the world is organized and orderly.

> Words form one of the basic order carriers of a child's world. Every joke, every departure from the norm, means an attempt to shake up this order himself or to hear it pushed around by another. But, in the final analysis, the order emerges unscathed. The child's relief that the order has proved unshakable is felt as happiness and breaks through as laughter [Quote taken from Alice Shabecoff, *New York Times Magazine*, September 15, 1968, p. 119].

Helmers felt that up to age 12, two-thirds of all laughter is of this type.

Much aggressive humor is funny, according to Helmers (1965), because the child is aware that it is only a fantasy situation. Whether a child is laughing because he knocked daddy down or because a cartoon character has been run over by a steam roller, it is funny (in the healthy child) only because he knows that each will get up again. If this did not occur, the child might be either confused or frightened, but the humor would likely be lost. While speech mistakes and other bumbling errors (e.g., slipping on a banana peel) are funny in their own right to the healthy child, they may become the source of cruel and derisive laughter in the child who feels unloved or unsure of himself.

Gestalt Theory

The only additional "classical" psychological theory or humor developed out of gestalt psychology. Maier (1932) laid the foundation for a gestalt theory of humor, adopting the traditional gestalt view that the meaning of an element or group of elements depends on the whole configuration of which it is a part. Changing some aspect of the whole may greatly change the meaning of both the whole and specific constituent elements. In a joke, a change in the anticipated meaning of particular elements leads to a sudden restructuring and change of interpretation of the whole. In a humorous experience, then, the suddenness and unexpectedness with which the restructuring occurs is the critical factor. Maier also felt that "objectivity" is important in the humor situation. That is, the humor can be maintained only if we do not identify with the figures involved, if salient emotions of our own do not come into play. However, since an objective attitude and abrupt restructurings also characterize productive thinking, other factors must uniquely typify humor. The differentiating factor "is found in the element of the ridiculous which is found in the humorous experience, but which is not present in reasoning [Maier, 1932, p. 72]." While the logic behind the humorous situation is unique to that situation, the logic behind productive thinking is consistent with the individual's general experience.

Bateson (1969) stressed the fact that a reversal of figure and ground occurs when we "get the point" of a joke. The structure of the joke draws our attention to certain elements while others are deemphasized, that is, form a background or setting for the joke. When the punch line occurs, the background material is suddenly and unexpectedly brought to the center of attention. Like Bateson, Fry (1963) suggested that the satisfaction derived from humor is due to the unexpected resolution of a series of paradoxes:

> During the unfolding of humor, one is suddenly confronted by an explicit-implicit reversal when the punch line is delivered . . . but the reversal also has the unique effect of forcing upon the humor participants an internal redefining of reality [p. 153].

Koestler (1964) has advanced a view highly consistent with the above, although he did not claim any gestalt leanings. He maintained that various types of creativity, humor, and scientific discovery are based on a characteristic mental process that he calls "bisociation"; that is, "the perceiving of a situation or idea . . . in two self-consistent but habitually incompatible frames of reference . . . [p. 35]." With respect to humor, bisociation

> causes an abrupt transfer of the train of thought from one matrix to another governed by a different logic or "rule of the game." But certain emotions,

> owing to their greater inertia and persistence, cannot follow such nimble jumps of thought; discarded by reason, they are worked off along channels of least resistance in laughter [p. 95].

None of the four theorists discussed above has attempted to extend the theoretical framework developed to children's humor. However, since each constitutes a primarily cognitive theory of humor, the role of children's level of cognitive functioning in the operation of the humor mechanisms described becomes an obvious area for investigation. If, as suggested by Maier, humor situations exhibit a unique kind of logic that does not generalize to other situations, it may be fruitful to investigate the role of acquisition of logical thinking in children's humor comprehension and appreciation.

Arousal Theory

More recently, two investigators interested in the more global problems of exploratory behavior (Berlyne, 1960, 1969) and the expression of positive and negative affect (Tomkins, 1962) have offered explanations of humor based on arousal factors operative in the humor situation. Berlyne (1960) suggested that when a humor stimulus is encountered, an orientation reaction occurs, representing an effort to understand the point of the humor depicted. He further maintained (1969) that "humorous situations always contain factors that can be expected to raise arousal and other factors that can be expected to lower arousal or else keep it within moderate bounds [p. 806]." The critical issue for this type of theory, then, becomes the identification of factors that (*a*) are capable of raising and lowering arousal; and (*b*) distinguish between the nature of the arousal process operating in humor, fear, startle, exploratory, problem solving, etc., responses. With respect to the former consideration, Berlyne rightfully pointed out that earlier philosophical theorists erred in emphasizing one or two arousal-influencing factors to the exclusion of all others, providing for theories with only limited explanatory power. He maintained that "all, or virtually all, of the factors that govern arousal can play a part in humor [p. 807]." The arousal-raising aspects of a humor situation, according to Berlyne, may be influenced either by the expression of some "emotional tendency" (sex, aggression, superiority, etc.), by the presence of collative variables in the stimulus situation (novelty, surprise, incongruity, etc.—all presumed to affect arousal through conflict), or by some combination of the two. The main ways in which arousal reduction may be brought about include gratification of the emotional tendency or motive or the achievement of an understanding of the "point" of the humor depicted. Of course, arousal reduction might also be achieved by more extrinsic means, such as shifting or withdrawal of attention.

Berlyne suggested that an "arousal jag" operates in many humor situations, while moderate increases in arousal operate in others. The arousal jag refers to

> a rise in arousal that is associated with a reward value, so that it is actively sought out, reported as "pleasant" or "enjoyable" by humor subjects, and capable of reinforcing instrumental responses that regularly bring it about [p. 806].

He cited a wide range of empirical evidence (Berlyne, 1960, 1967), pointing to the reward value of both arousal increases followed by decreases, and moderate increases in arousal, suggesting that these may operate either independently or together in humor situations. With respect to the second issue specified earlier, however, no explanation has yet been offered for the differential nature of the operation of the arousal jag or "boost" in humorous and nonhumorous situations. Until such a distinction has been accomplished, this arousal theory must be considered an incomplete explanation of humor.

Tomkins (1962) has advanced a general theory of the innate activators of positive and negative affect. The critical differentiating dimension here is the density of neural firing or stimulation; that is, "the product of the intensity times the number of neural firings per unit time [p. 251]." While both positive and negative affects (e.g., interest, fear, and startle, which differ only in rate of stimulation increase) may be produced by sudden increases in stimulation, Tompkins maintained that only positive affects (e.g., laughter and joy) are activated by stimulation decrease. He considered laughter to be a more intense form of the smile, and:

> As such it is activated in the same way as the smile except that the general density of stimulation which is suddenly reduced begins at a higher level in the case of laughter compared with the smile [p. 370].

The quality of laughter or smile produced, according to this view, depends on the nature of the originally more intense stimulation. In any case, a certain minimal steepness of the gradient of reduction must be reached before smiling or laughing may occur, with the suddenness of the drop and the original intensity level interacting to determine the nature of the response.

While the language used by Berlyne (1960, 1969) and Tomkins (1962) differs, Berlyne's use of the concept of arousal increase and decrease is comparable to Tomkins' use of the concept of increase and decrease of density of stimulation. There would appear to be agreement that a decrease in this dimension, or an increase followed by a decrease, plays an important role in the production of smiling or laughter. For Tomkins, a certain optimal level of this dimension is both necessary and sufficient for the production of such a response; for Berlyne, however, it is neither necessary nor sufficient. It is not necessary since Berlyne maintained that moderate rises in arousal may be conducive to a humor response apart from any consequent arousal reduction. It is not sufficient since the arousal jag is postulated to operate in similar fashion in nonhumorous situations.

While Berlyne (1960, 1969) has not directed his theory to developmental considerations, Tomkins (1962) reviewed the literature on the development of the smiling response in infancy, and interpreted the data discussed according to the operation of the mechanism of reduction in density of stimulation described above. He suggested that both the smiling response and its triggering mechanism are inherited, and he appears to maintain that the nature of the operation of this mechanism does not change as the child develops. However, he concluded that learning and memory determine which stimuli initiate the mechanism. An inherited receptivity of the mechanism to particular stimuli (i.e., "releasers") is rejected as very unlikely.

Kagan (1967), in discussing the development of the smiling response to the human face, suggested that "the creation of a schema for an event is one major source of pleasure [p. 135]," where schema means the "representation of an external pattern." According to Kagan, when a 4-month-old infant smiles to a photograph of a regular face, he has actually made a cognitive discovery. "This discovery occurs when the infant is able to match stimulus to schema. A moderate amount of mismatch between event and schema is necessary as a source of pleasure. He further argued:

> Sixteen-week-old infants and eight-year-old children smile spontaneously at events that seem to have one thing in common—the event is a partial match to an existing schema and an active process of recognitory assimilation must occur [p. 137].

More recently, Kagan[3] advanced the view that the critical factor in this process is the gradual buildup and release of tension as the child exerts mental effort to assimilate the new situation into old schemas. Thus, if assimilation is either immediate or so difficult that no effort is made toward assimilation, there will be no buildup of tension and, consequently, no smile. Kagan's postulation of tension reduction as the basis for smiling clearly bears a similarity to the shift in arousal and density of stimulation suggested by Berlyne (1960, 1969) and Tomkins (1962). Kagan's position is a valuable addition, however, in that, like Berlyne, he related the arousal fluctuation, which gives rise to smiling, to both the present stimulus situation and the child's past experience. He went beyond Berlyne and Tomkins, however, in describing the cognitive process involved in producing the smile. That is, a smile is likely to occur when the child is initially unable to incorporate an object or situation into his existing schema system, but does so after exerting a more active assimilatory effort.

Developmental Theory of Cognitive Humor

The theoretical position advanced below was developed on the assumption that in light of the complexity of the humor process, it might be fruitful to consider the development of different types of humor separately. An important distinction in this respect may be drawn between humor based primarily upon the gratification of some salient need (often referred to as "tendential" humor) and humor based primarily on a violation of expectations the individual has come to hold for relationships among different aspects of his environment. The term "cognitive" humor is used here for the latter type, since it is assumed that appreciation of such humor is primarily an intellectual phenomenon, with satisfaction of emotional needs contributing minimally. This type of humor is typically experienced as funny in the sense of being "clever," and is generally accompanied by milder forms of affective expression. Thus, the well-known cartoon of a skier whose skis have left tracks on each side of a tree generates humor only because of the obvious impossibility of the feat in light of our prior experience with the stimulus elements depicted. Sexual, aggressive, etc., elements appear to be irrelevant to the humor depicted here.

---

[3] Kagan, J. On continuity: A demonstration from infancy. Paper presented at the meeting of the Society for Research in Child Development, Santa Monica, California, March 1969.

The writer agrees with Kagan's (1967) view that the assimilation of a stimulus into an already established schema constitutes a source of pleasure in the infant, where some optimal moderate amount of mismatch between stimulus and schema occurs. Berlyne (1960, 1967) has demonstrated that such "novel" stimuli lead to increases in arousal. While moderate increases in arousal have some reward value in their own right (Berlyne, 1967), the primary source of pleasure here is probably due to the arousal decrease accompanying the infant's successful effort to assimilate the stimulus into the relevant schema. Since one of the earliest perceptual schemas developed by most infants is that of the human face, it is used to exemplify the writer's theoretical position. Kagan (1967) has demonstrated that the pleasure assumed to derive from the operation of the mechanism described above tends to be accompanied by smiling in the infant. Kagan showed various patterns representative of human faces to the same group of infants at 4, 8, and 13 months of age. While significantly more frequent smiling occurred to a photograph or sculpture of a regular face than to a regular schematic face or disarranged face at 4 months, by 8 months of age equal amounts of smiling were shown to the regular and distorted faces. By 13 months of age, most smiling occurred to such alterations of the human form as a three-headed man and a mule's head on a man's body. The writer would agree with Kagan's explanation that these developmental changes in smiling reflect the ease with which the stimuli could be assimilated into the infant's facial or human-body schemas. Maximal smiling to increasingly greater distortions of a regular face with increasing age indicate that "making sense" out of a regular or only slightly distorted face becomes so easy that it generates little arousal, and consequently, little pleasure.

The critical issue here lies in the determination of the point at which the above source of pleasure develops into humor. While a smile or laugh may be evoked in the above situation, it does not become reflective of a humor experience until conceptual capacities have been acquired. Prior to the onset of conceptual thinking, the child experiences expectancy violation (i.e., a stimulus requiring some accommodation for successful assimilation) only in the sense of the stimulus being different or novel. The basic prerequisite for experiencing such expectancy violations as humorous, however, lies in the ability to perceive the stimulus not only as being different but also as *inconsistent* with the child's prior experience with the stimulus situation. Thus, the pleasure in being able to incorporate new stimuli into existing structures becomes transformed into the pleasure of humor when the assimilatory process is accompanied by the realization that the stimulus situation is "backwards," "wrong," or "not like it should be." A "reality–fantasy" dimension becomes important here, in that the child perceives expectancy violations as being funny only when he has acquired a stable enough conceptual grasp of "the real world" that he can assimilate the disconfirmed expectancy as being only a play on reality (see McGhee, in press). Thus, the disarranged face or three-headed man in Kagan's (1967) experiments becomes funny only when the child realizes that what is depicted could not really occur, except at a fantasy level. Prior to the achievement of such conceptual mastery over the environment, an effort will be made to incorporate the new stimulus into relevant schemas the child has developed; that is, "reality assimilation" will occur. With higher levels of conceptual mastery, however, "fantasy assimilation" occurs. While the child does assimilate the source of the expectancy disconfirmation into his existing cognitive structures, accommodation does not occur. The new situation is assimilated only in a fantasy or pretend fashion, and it is the comparison of the normal criteria governing complete assimilation into a schema with the present violations of those criteria that causes the child to see the situation as funny. While this type of fantasy assimilation generates pleasure in the same fashion as described above for the normal incorporative functioning of the preconceptual infant, an additional source of pleasure derives from the child's certainty that the stimulus situation depicted does not really exist. As long as the child remains in a reality mode of assimilation, expectancy disconfirmations may generate interest, confusion, or fear, but will not lead to humor. Moreover, as long as the reality orientation is maintained, the child may have great difficulty in assimilating the new situation. By switching to a fantasy mode of assimilation, the child is immediately able to make sense out of the stimulus and construes it as a play on reality. The pleasure derived from his success at identifying the nature of the reality violation is manifest in the form of laughter.

The child's ability to identify cues indicating which mode of assimilation is appropriate plays an important role here. That is, when a cartoon or joke is presented to the child, he knows that the elements depicted only occur at a fantasy level and that to successfully understand the "point" of the joke, he must understand the way in which reality factors have been violated. If the child does not perceive the cues indicating a potential humor situation, or perceives cues suggesting the opposite, he will attempt to assimilate the new material into his cognitive structures in normal fashion and will fail to perceive any humor in the situation.

While the conceptual child is able to see humor in the inconsistencies he identifies in jokes, cartoons, word pronunciations, object labels, etc., logic plays no part in the process. Thus, the child laughs at the man with a mule's head simply because it is wrong, or discrepant from his prior experience with the stimulus elements involved. One reason young children have difficulty explaining the basis for humor in such cases (along with their limited language development) is that they are unable to go beyond reproducing a memory store of the way the situation has occurred in the past. Only with the acquisition of the capacity for logical thinking (this should develop around age 7, according to Piaget, 1950) is the child able to consider the violation of logical relationships among stimulus elements as a new source of humor. Thus, the same humor stimulus may be laughed at by prelogical and logical children, but the latter have acquired a "higher level" of appreciation. Along with providing for a better understanding of humor situations that younger children are also capable of responding to, the acquisition of logical thought capacity also allows the child to comprehend more abstract sources of humor. While for the younger child, expectancy violations must take the form of some type of discrepancy of sensory input relative to his previous experience, the child who has acquired logical thinking may find humor in the logical implications of a configuration of elements that are perfectly consistent with the child's prior sensory experience. It is the acquisition of this capacity that carries the child to the level of humor responsiveness found in adults.

EMPIRICAL STUDIES

Among those studies that have used children as subjects, few have attempted delineate developmental changes in children's response to humor stimuli. Even fewer have related their findings to any type of theoretical framework.

Age Differences

The most typical approach to studying children's humor has been to simply note age differences in children's response to different types of humor situations. In an early study, Walker and Washburn (1919) presented part of a picture to fourth grade, seventh grade, and college subjects and asked the subject to first complete the picture with one of three alternatives and then rate the funniness of each alternative completion. While one alternative formed a normal or appropriate situation, the other two formed incongruous situations. For both types of completions, fourth graders gave higher mean ratings than seventh graders, who in turn gave higher ratings that college subjects. In an effort to test several theoretical views on humor, Justin (1932) observed the amount of laughter of boys and girls at ages 3, 4, 5, and 6 in the following types of situations: surprise or defeated expectation, superiority and degradation, incongruity and contrast, social smile as a stimulus, relief from strain, and play situations. In general, she found an increase in humor responsiveness up to age 5, and a decrease in the sixth year. While incongruity was found to increase in its effectiveness at producing laughter as a function of age, no age differences were found for the other humor categories observed. For all ages tested, however, children showed more laughter in situations in which they actually participated. Finally, for 3–5-year-old children, verbal presentation was found to be more effective than pictorial presentation. Ding and Jersild (1932) found no significant age differences in laughter or smiling of 2–5-year-old children in a nursery school situation.

In a study dealing more with humor comprehension than appreciation, Schaffer (1930) asked subjects in Grades 4–12 to explain the meaning of 10 caricature-type cartoons dealing with important social, economic, or political problems. Using a nationwide sample, he found that subjects in Grades 4–5 gave primarily either descriptive responses or concrete interpretations of the cartoons. After Grade 8, however, abstract interpretations constituted the most frequent response. Wells (1934) asked subjects in Grades 7, 9, and 12 to order four different types of humorous literature in terms of preference for the type of humor depicted. For the group as a whole, the following order of preference (from greatest to least) was found: (*a*) absurdity, (*b*) slapstick, (*c*) satire, (*d*) whimsy. While taste for absurdity decreased slightly, preference for satire and whimsy showed a slight increase between Grades 7 and 12. Little change with grade level was noted for slapstick. Wells noted that in the popular absurdity literature, it was the incongruity and unexpectedness of situations that subjects found amusing. Brumbaugh (1940) had trained observers note all the incidents of laughter in the classrooms of 1,600 Grade 3–6 boys and girls. She found that at all four grade levels, fantastic stories, such as fairy tales, elicited more amusement than any other type of story.

Using 709 children between 7 and 18 years of age in Scotland, Laing (1939) asked subjects to tell their funniest jokes, their most amusing experience, describe the most comical people they have seen, etc. Breaking his subjects into three age groups (7–10, 11–13, 14–18), he found the humor of both younger groups to be predominantly visual; these subjects showed little appreciation of verbal wit. Incidents involving someone's discomfiture were given most frequently by the middle group. The adolescents of the oldest group were distinguished most by their (*a*) marked individual differences in sense of humor, (*b*) greater appreciation of verbal wit, and (*c*) tendency to reflect on why they found particular situations funny.

Behar and Bevan (1956), although not specifically interested in humor, studied the perception of incongruity in young children. Children ranging from 4 to 7 years of age were shown a photograph of a three-sided building and were asked to describe what they saw. The photograph was taken from such an angle that, unlike the surrounding buildings, the three-sided building appeared to have only a front wall—even though it did look inhabited. They found little recognition of the apparent incongruity of this situation by age 4. While a rapid increase in recognition was found between 5½ and 6½, there was still very little spontaneous recognition of incongruity by age 7. Graham (1958) presented 44 cartoons to first, third, and fifth graders in sets of four, and asked the subject to choose the funniest of the four. His data showed significant differences in the three grade levels. He suggested on the basis of his data that "additional experience and/or maturation . . . increases the number of objects or situations which will be perceived as humorous [p. 327]." Subjects at all three grade levels were also found to be equally unable to explain why the cartoons were funny. Kreitler and Kreitler (1970) asked ninety-two 5–6-year-old children of European and Oriental descent to tell what they thought about 15 pictures depicting absurd situations. They found that the cognitive strategies that facilitated laughter and smiling at an above-chance level involved identification of the absurd nature of the situation, with no accompanying effort to resolve the absurdity.

Zigler, Levine, and Gould (1966a) showed a series of cartoons to subjects in Grades 2, 3, 4, and 5, making an effort to use difficulty levels and content "appropriate" for children. They found a significant positive relationship between grade level and comprehension of the cartoons. Using smiling and laughter as an index of humor appreciation, the "mirth" response was found to increase steadily between Grades 2 and 4, but decreased significantly between Grades 4 and 5. McGhee (1971a) showed 5-, 7-, and 9-year-old boys cartoons and jokes representing a primarily "cognitive" type of humor (i.e., emotional or tendential bases for humor were eliminated to as great an extent as possible) and found that comprehension of all humor stimuli used increased significantly with age. No consistent relationship was found, however, between age and humor appreciation. For more complex classes of humor stimuli, older subjects also showed significantly more frequent interpretive (as opposed to descriptive) explanations of the basis for humor depicted. Using the same set of subjects, McGhee (1971b) used three different types of humor tasks, requiring the subject to (*a*) arrange sequential cartoons in the funniest order, (*b*) choose the funniest pictorial completion of a cartoon, or (*c*) choose the funniest caption. Older subjects demonstrated significantly greater comprehension and more frequent "correct" choices than younger subjects for all three humor tasks. Older subjects also showed a significantly greater number of interpretive humor explanations for the latter two humor categories, and greater humor appreciation for cartoons in the difficult caption-matching task.

The wide variability in both age of subjects and operational measures used to collect humor data in these studies prohibits us from drawing any firm conclusions about age differences in children's humor. There is some suggestion, however, that younger children prefer a more direct kind of visual or motor involvement with potential humor situations, while older children are able to use their developing thought capacities to identify and appreciate more abstract sources of humor. While the greater knowledge and more advanced cognitive capacities of older children do provide for greater comprehension of humor stimuli, they do not necessarily lead the child to perceive them as funnier. Complexity of the humor stimulus would certainly appear to be a factor here and is discussed in a later section.

Preschool Children's Humor

All of the studies of preschool children's humor known to the writer were carried out over three decades ago and used simple observational techniques, typically in a nursery school setting. Jones (1926) observed nine children between 16 and 36 months of age in a nursery home. She found the most frequent laughter-evoking situations for this group to be tickling, playfully dressing, gentle bathing, romping with other children, and teasing. She also observed that the same stimuli might arouse laughter on one occasion, but crying on another. Enders (1927) observed 2–5-year-old children's laughter in a nursery school during free play periods and upon presentation of toys, pictures, false faces, and games involving vigorous arm and leg movements. The situations most effective in producing laughter were those involving changes in sound and/or movement. Gregg[4] observed 22 nursery school children for 40 hours and found that laughter occurred most frequently in connection with some type of physical activity. Washburn (1929) studied the laughter of a group of infants monthly between the first and twelfth month of age and found no significant relationship between amount of laughter and chronological age, developmental level, or a height-weight index. Buhler (1930) also studied infants' laughter. She found that laughing usually begins by the second month and is initially evoked by another person's glance or voice up to the fifth month. Between the fifth and ninth month, the infant shows a progression of laughter in the following situations: accompanying play movements (5 months), upon receiving bodily stimulation (6 months), when offered toys (7 months), upon hearing words (8 months), upon changes in posture (8 months), and upon awakening (8 months).

---

[4] Gregg, A. An observational study of humor in three year olds. Unpublished master's thesis, Columbia University, 1928.

Kenderdine (1931) recorded 2–4-year-old children's spontaneous laughter in a nursery school setting and upon presentation of toys in an experimental situation. Few children showed laughter upon presentation of the toys; rather, most children showed interest or curiosity. The two situations found to induce laughter most frequently in these children were those that (*a*) they realized were socially unacceptable and (*b*) involved bodily motion. Ding and Jersild (1932) observed the smiling and laughter of fifty-nine 2–5-year-old Chinese children during free play and normal activities in a combined kindergarten and nursery school. They found that both smiling and laughter occurred most frequently in association with general motor activity. Also, children who smiled most often showed the greatest amount of laughter. They suggested that laughing and smiling serve as a motor outlet in response to many forms of stimulation in young children. Brackett (1933) observed twenty-nine 18–48-month-old children in a nursery school free play situation and during routine adult-directed activities. She found a high degree of consistency in children's laughter in the two types of situations. Also, subjects ranking high in laughter in this study ranked highest one year later. Using the same set of subjects, Brackett (1934) found positive relationships between frequency of laughter and chronological age (.44), mental age (.26), and a height-weight index (.89).

The data from these studies consistently point to the importance of direct motor involvement in the laughter of very young children. The most typical occurrences of laughter, especially in free play situations, appear to accompany engagement in some ongoing activity and to reflect the child's enjoyment in exercising the motor skills he has developed. While this laughter may reflect a basis for humor in the degree of mastery of various motor skills relative to other children, in the fashion suggested by some psychoanalytic theorists, it is equally likely that such instances of laughter are not indicative of humor in the sense of something being funny. The laughter may simply be reflective of pleasure derived from the incorporation of a wide range of environmental stimuli or tasks into the motor skills developed.

Cognitive Development and Humor

While studies demonstrating age differences in humor comprehension or appreciation do shed valuable light upon changes in the humor process as the child develops, age in itself is not an explanatory variable. Consequently, a number of investigators have gone beyond a simple description of age differences, attempting to account for such differences in terms of differential level of cognitive development.

Bird (1925) used elementary school subjects and found a correlation of .89 between IQ and success in a humor test based on absurdity situations. Enders (1927) and Ding and Jersild (1932) failed to find any significant relationship between measured IQ and amount of laughter exhibited by 2–5-year-olds in a nursery school setting. Kenderdine (1931) also observed nursery school children's laughter and found that in a situation with 100 possibilities for laughter, a group with a mean IQ of 141 averaged 13 instances of laughter, as opposed to 4.7 for a group with a mean IQ of 118. Justin (1932) found correlations between IQ and number of laughter responses to an incongruity situation to range from .40 at age 3 to .12 at age 6. Williams (1946) had 11- and 12-year-olds rank cartoons and jokes from most to least funny. Defining humor appreciation in terms of degree of agreement with a standard order obtained from adults, he found greater appreciation with more intelligent children. Overlade[5] found a significant positive relationship between comprehension of jokes and American Council of Education test performance, the ability to think abstractly, and the ability to discern embedded figures.

---

[5] Overlade, T. Humor appreciation as abstract ability. Unpublished doctoral dissertation, Purdue University, 1954.

The data obtained in these studies are difficult to interpret both because of the relative lack of control exercised over the nature of the stimulus material used and the frequent failure to distinguish between comprehension and appreciation of humor. With respect to the first issue, although a wide range of ages of subjects has been used, in many cases no effort was made to control for complexity or difficulty of the humor items for the particular subjects sampled. If a given series of cartoons or jokes is either extremely simple or very complex for the age group in question, it is unlikely that either comprehension or appreciation would correlate highly with IQ. The role of IQ in each of these variables should be manifest most clearly in cartoons or jokes at a moderate level of difficulty.

Concerning the second issue, some of the above studies dealt with comprehension while others were primarily concerned with appreciation. In many cases, however, it is difficult to determine the extent to which either variable is contributing to the data obtained. It would seem to follow, up to a point at least, that the greater the child's understanding of a joke or cartoon, the greater should be his appreciation of it. But, as pointed out by Zigler et al. (1966a), a cartoon or joke that is either very easily understood or incomprehensible to the child is unlikely to be perceived as funny. If the ease with which a child comprehends a humor stimulus does play an important role in determining his appreciation of it, then we clearly cannot draw any inferences about perceived funniness on the basis of degree of comprehension. Similarly, it should be difficult to predict comprehension given perceived funniness; for the child may be responding to some aspect of the stimulus other than that "intended" by the experimenter. Thus, the child who laughs at a given situation may understand it very well, or not at all. His response may be based on anything from purely perceptual aspects of the stimulus to abstract implications of the events depicted. Of course, dimensions other than these purely cognitive ones will generally also play an important role in determining perceived funniness (e.g., themes of aggression, superiority, etc.). In short, given that we have information on only one of these variables, we have learned very little about the nature of the child's humor response.

In a more recent study, Zigler et al. (1966a) did investigate the relationship between humor comprehension and appreciation. Using subjects in Grades 2, 3, 4, and 5, they defined cognitive development in terms of age level, assuming level of complexity of cognitive functioning to increase progressively with age. As noted earlier, their findings suggested a curvilinear relationship between these two variables. While comprehension consistently increased with age, appreciation increased through Grade 4, but then declined significantly by Grade 5. In addition, within each grade level, subjects showing higher comprehension scores also showed greater mean mirth response scores. Zigler and his associates advanced a "cognitive congruency" principle to explain this decrease in appreciation with age, arguing that "cartoons which make few cognitive demands elicit a lower mirth response than those that are in keeping with the complexity of the child's cognitive apparatus [p. 514]." In order to test this principle, Zigler, Levine, and Gould (1967) completed a follow-up study in which an effort was made to control for the level of cognitive demands made on the subject by a given humor stimulus. Groups of cartoons were matched in terms of difficulty level for each of three grade levels (3, 5, and 7). That is, on the basis of performance of another group of subjects on a larger sample of cartoons, the experimenter selected four groups (representing four difficulty levels) of cartoons for each grade level, such that a given group of cartoons for one grade was equally difficult (as judged by comprehension scores) for subjects in that grade as the corresponding group for subjects in another grade. Using overt expression of affect and cartoon preference scores as measures of appreciation, their data showed a peak in humor appreciation for the moderately difficult cartoons in all three grades. Zigler et al. (1967) concluded that these findings support the cognitive congruency principle, since

> The magnitude of the mirth response was found to depend upon the degree of cognitive congruence existing between the cognitive demand features of the humor stimulus and the cognitive resources of the individual [p. 335].

As pointed out by McGhee (1971a), however, in defining level of cognitive functioning in terms of comprehension scores, Zigler et al. (1967) did not attempt to specify either (*a*) systematic differences in cognitive resources possessed by subjects obtaining different comprehension scores or (*b*) how the cognitive demand features vary among their different difficulty levels.

McGhee (1971a) used a Piagetian theoretical framework in an effort to overcome these two deficiencies. The level of cognitive development of 5-, 7-, and 9-year-old boys was measured in terms of performance on four different Piaget tasks. This study investigated the importance of degree of acquisition of operational thinking (within each grade level) for the comprehension and appreciation of two distinct types of cartoons and jokes with a primarily cognitive basis for humor (as defined earlier). These two humor categories, labeled novelty and incongruity humor, differed most essentially in the level of cognitive processing hypothesized to be necessary for their comprehension. For stimuli in each category, the basis for humor consisted of violations of expectations that the subject has presumably come to hold regarding the contextual composition of various physical or behavioral events in the stimulus situation. In novelty humor, the expectancy violation was manifest in the form of something visually or physically discrepant from the subject's prior experience. In incongruity humor, however, nothing was discrepant about the physical organization of elements portrayed. Expectancy violation for cartoons and jokes in this category only occurred at a more abstract level upon assimilation of information provided by different cues in the stimulus. McGhee was testing the hypothesis that for 7-year-olds (who are in transition between preoperational and operational modes of functioning), degree of acquisition of logical thinking is positively related to the capacity to identify the basis for humor in incongruity stimuli. For the more simple novelty humor, it was hypothesized that logical thinking should not be necessary for comprehension, since expectancy violations constituting the basis for humor occur at a visual level. No significant relationships were predicted for 5- and 9-year-olds, since most 5-year-olds have not yet begun to acquire operational thinking, while most 9-year-olds are well along in this development.

The data were clearly supportive of this hypothesis. For the two classes of novelty humor in the 7-year-old group, degree of acquisition of operational thinking was not significantly related to humor comprehension. Significant positive findings were obtained, however, for all three classes of incongruity humor. More cognitively advanced 7-year-olds also gave significantly more frequent interpretive (as opposed to descriptive) explanations of the basis for humor for incongruity humor, but not for novelty humor. As hypothesized, none of these relationships was significant at the 5- and 9-year age levels. Analyses failed to show any significant linear or curvilinear relationship at any age level between cognitive development and humor appreciation (defined in terms of the subjects' ratings on a 1–5 funniness scale). McGhee concluded that no support was obtained for the cognitive congruency principle.

McGhee (1971b) used the same subjects used in the previous study (ages 5, 7, and 9) in a different approach to studying the role of operational structures in determining comprehension and appreciation of humor. Three different sets of humor stimuli, containing varying degrees of emphasis on tendential versus cognitive bases for humor, were used: (*a*) sequential cartoons, in which the subject must arrange the parts in the funniest order (Set I); (*b*) cartoons in which some part is missing, and the subject must pick one of three alternative completions that makes the cartoon the funniest (Set II); (*c*) cartoons for which the subject must choose the funniest of three possible captions (Set III). Significant positive relationships were found between degree of acquisition of operational thinking and humor comprehension for Sets I and II with 7-year-olds, and for

Set III with 9-year-olds. While more cognitively advanced subjects consistently made more frequent "correct" orderings or choices at all age levels, this relationship only reached significance for Sets I and II at the 7- and 9-year level. For the age 7 group, a significant relationship was found between level of operational thinking and the tendency to give interpretive humor explanations for Sets I and III. Again, no significant linear or curvilinear relationships were found between humor appreciation and level of cognitive development at any age level. While no significant curvilinear trends between humor comprehension and appreciation were present at any age level, a significant positive linear trend was found for Set II at age 9.

In general, the data presented in this section suggest that more cognitively advanced children are likely to demonstrate better comprehension of most types of humor stimuli. However, they do not necessarily show a corresponding increase in appreciation of such humor. It is unclear at this point how a child's understanding of a joke, cartoon, or real-life situation corresponds to how much he laughs at it or how funny he says it is. While earlier studies tended not to consider the distinction between comprehension and appreciation of humor situations, it is clearly an important and meaningful one. It would seem imperative that all future research endeavors in this area measure the child's status along each of these dimensions if we are to expand our understanding of the processes operative in children's humor. It is also clear that future research must examine closely the nature of the stimulus materials used. In much of the research reviewed here, the primary consideration in choice of stimulus materials has been that they be "appropriate" for children. In most cases, this has meant the minimizing or elimination of overt expression of sexual or aggressive themes. While stimuli have been categorized along particular dimensions (e.g., content area), there has seldom been any other effort to consider additional dimensions along which the stimuli might vary. In most areas of research, a great deal of effort is made to clearly define the stimulus situation. In research on humor in adults as well as children, however, this consideration has in many cases been paradoxically overlooked.

Although all of the studies reviewed in this section investigated the role of level of cognitive functioning in children's humor, the approaches to defining this variable have varied considerably. The simplest approach, of course, is to merely use age and assume that level of cognitive development increases progressively with age. However, since widespread individual differences accompany most developmental phenomena, if we are to attribute a child's performance to a given cognitive dimension, some measure of that dimension should be obtained.

The most typical approach to conceptualizing level of cognitive functioning has been in terms of various IQ or related measures. While the relatively consistent positive findings using this approach with children do support its usefulness, the use of IQ scores alone does not clarify the role of acquisition of particular types of cognitive skills for understanding, appreciating, or generating different types of humor. Furthermore, McGhee (1971a) found that the degree of acquisition of concrete operational thinking in 7-year-old children accounted for variance that could not be accounted for in terms of IQ scores. Thus, while measured IQ may be tapping cognitive skills that are very important in the humor process, the developing child is also acquiring other new cognitive capacities that are not measured by conventional IQ tests, but play an important role in the comprehension of at least certain types of humor.

Social Nature of Humor Situations

While it is generally assumed that laughter and humor are primarily social phenomena, the amount of data on children relating to this issue is surprisingly limited. Most of these studies were completed nearly four decades ago, and all were confined to children 5 years of age or younger, using the observational approach in nursery school settings. Enders (1927), Gregg (see Footnote 4), Kenderdine (1931), Ding and Jersild (1932), Brackett (1933, 1934) all found more frequent laughter in nursery-school-age children in social situations than when alone. While Brackett

(1933) found frequent laughter in the presence of adults as well as children, Enders found that children seldom laughed in the presence of adults. Ding and Jersild found more laughter and smiling in the company of close friends, brothers, or sisters than with strangers. Brackett (1933, 1934) found that children ranking high in amount of laughter exhibited tended to prefer other members of the group showing the same behavior. Also, most children were found to confine most of their laughter to a few particular children.

To the writer's knowledge, no systematic attempt has been made to determine the nature of those stimulus situations that tend to evoke laughter when the child is alone. Valentine (1942), however, after reviewing the literature and making personal observations of his own, did discover a number of occasions for children's laughter in nonsocial situations. These included sight of a bright or pleasing object, jogging, mild shock or surprise, repetition, the incongruous, recognition, accomplishment of some new task or activity, incongruity in words or ideas, and coincidental events.

While the available data do suggest the importance of social factors in much of children's laughter, they do not foster the conclusion that laughter is primarily a social phenomenon, to the exclusion of nonsocial factors. Since the nursery school environment is very conducive to social interaction in general, it follows that a large proportion of laughter occurring in this setting will be social in nature. Similarly, since the stimulus evoking the laughter is likely to be shared by others with whom the child may be interacting, the laughing child is likely to be present in a situation where other children are also laughing. This type of experimental situation, therefore, does not provide suitable conditions for determining the nature of the evoking conditions or the relative importance of laughter in nonsocial settings.

An important distinction must be made again here between laughter as an index of humor and as a mere expression of heightened pleasure. Of course, the child may also laugh simply to avoid giving the appearance of not understanding or to maintain social desirability. Thus, the critical issue is (*a*) whether the smile or laugh necessarily reflects a humor experience, and (*b*) whether a social environment—with and/or without social laughter—is capable of either creating humor out of a previously nonhumorous situation or boosting the perceived funniness of an already humorous situation.

Language and Children's Humor

While language would appear to be central in much of children's humor, there has been little systematic investigation of the role it does play and how this changes with age. However, a few writers have theorized about its importance on the basis of observation and anecdotal evidence. Bever (1968) found that 5-year-olds laughed more at jokes involving ambiguities in grammatical surface structure (e.g., "Why can't you starve in the desert?" "Because of the sand-witches there"), while 10-year-olds laughed more at jokes involving ambiguities in deep structure (e.g., "Why can you jump higher than the Empire State Building?" "Because it can't jump at all"). Gesell and Ilg (1946) argued that by the third year the verbal play of repeating a word (like "golly") with someone becomes humorous. By the fourth year, he begins to enjoy silly rhyming (e.g., "mitsy, witsy, bitsy") and play on words. Wolfenstein (1954) suggested that the earliest form of the young child's word play centers around the ambiguity of his own sexual identity. Generally by the age of 3, the child has begun to master any doubts about his own sex and can accept "you're a boy (or girl)" as a joke. But prior to this, there is a tendency to take the statement seriously and be threatened, for sexual identity has not yet been fully established. The mastery of individual identity sets the stage for the next level of verbal humor about the fourth year. Again, being called by the wrong name may be disturbing prior to this mastery, but funny thereafter. Changes in common names generally come several years later than changes in proper names, with the latter continuing to increase in frequency up to adolescence. While the jokes or humor of the first 5 or 6 years tend to be spontaneous and of the child's own invention, the child relies on a concrete verbal formula between about 6 and 11. Children memorize jokes in this period, seldom making up their own. Grotjahn (1957) advanced a view similar to that of Wolfenstein (1954), arguing that children begin to see the humor in jokes when they have mastered the language characteristics utilized by the joke.

Helmers (1965) argued that while the child begins to play with sound changes by age 2, he does not begin laughing at funny speech patterns until he has achieved a relatively high degree of mastery over language in the third or fourth year. While the 3-year-old makes word mistakes accidentally and laughs only when he realizes the mistake, the 4-year-old produces verbal humor by changing word arrangements or sounds on purpose. It is only by age 5, according to Helmers, that children acquire the capacity to understand jokes. Once the child has mastered language, he is likely to find any change in normal speech to be funny, especially if it is due to the accidental errors in speech made by a younger child.

While the paucity of empirical data again restricts the conclusions that may be drawn here, there is fairly consistent agreement that degree of language mastery plays a key role in the manner and extent to which language becomes a source of humor. While there are no available data to resolve the issue, it appears likely that very young children begin to laugh at changes in familiar word sounds before conceptual thought capacities have developed. Once new concepts begin to be acquired, they appear to become occasions for laughter only when the child has acquired some competency in their meaning and usage. Modification of a recently acquired concept may be funny if initiated by the child himself but disconcerting if initiated by another. Since the child is still learning about the range of events encompassed by the new concept, he encounters a conflict of whether to take the new information seriously or as a joke. As he gains a more firm conceptual grasp of the concept in question, any deviation is more likely to be placed in the joke category. An important factor here may lie in the child's growing capacity to identify cues given by the speaker that indicate whether the inconsistency (with the child's past experience) is put forth in a serious or joking manner.

A central issue here again lies in the question of when laughter as a simple manifestation of pleasure becomes indicative of the experience of humor, for early alterations of word sounds may lead to pleasure solely on the basis of successful assimilation into "word schemas" in the fashion described earlier. It may be that alterations of word sounds or meanings become a basis for humor only after conceptual thinking has developed, permitting the child to realize the inconsistency depicted.

Sex Differences

While a number of studies using boys and girls as subjects have failed to discuss the presence of sex differences, those studies that have systematically considered sex differences have yielded inconsistent data. Kimmins (1928) observed that girls show a greater appreciation of the ludicrous at earlier ages, while boys show more at later ages. Brumbaugh[6] noted that boys produce more original humorous drawings than girls, although individual differences within each sex were great enough to make the distinction questionable. Justin (1932) subjected 3–6-year-old children to a number of different types of laughter-provoking situations, but found no marked sex differences. However, girls were found to smile more than boys, while boys tended to laugh more than girls. Boys also showed more variability in their responsiveness. Williams (1946) found no sex differences in degree of appreciation of jokes and cartoons among 11- and 12-year-olds, but boys again showed more variability in appreciation scores. Zigler et al. (1966a, 1966b, 1967) failed to find any significant sex differences in comprehension or appreciation of cartoons among either normal or retarded subjects.

Although the data in this area are again relatively meager, the solid experimental approach of the Zigler et al. studies suggests that through the ninth or tenth year, at least, the different social environment in which boys and girls mature does not lead to significant differences in their ability to understand or appreciate cartoons. Justin's data, however, are partially conflicting here, since she found boys to be more overtly expressive of their appreciation. Whether this does in fact indicate the development of a more "reserved" humor responsiveness in females than in males can only be determined by further research on this issue.

---

[6] Brumbaugh, F. Stimuli which cause laughter in children. Unpublished doctoral dissertation, New York University, 1939.

Humor in Exceptional Children

To the reviewer's knowledge, only two published studies have attempted to study the humor of physically or intellectually handicapped children. In light of the theoretical view that appreciation of humor is the result of the gratification of repressed (Freud, 1916) or suppressed (Strickland, 1959) wishes, Hetherington (1964) reasoned that physically handicapped children, whose ability to participate in physical activities is extremely limited, might satisfy their suppressed desire for physical activity through the expression of humor. Thus, she investigated the relationship between motor disabilities and humor ratings given to cartoons portraying activity or nonactivity by 10–12-year-old poliomyelitis, cerebral palsy, and normal children. The poliomyelitis children rated activity cartoons as significantly more funny than nonactivity cartoons, and also gave higher ratings to activity cartoons than did cerebral palsy or normal children. Thus, support was obtained for psychoanalytic theory.

Zigler et al. (1966b) completed a study to determine "whether the restricted experiences and/or unique personality dynamics of the retarded influenced the humor response [p. 473]." They attempted to control for general cognitive level by using normals, institutionalized retardates, and noninstitutionalized retardates matched for mental age. Twenty subjects (10 males and 10 females) in each category were shown 25 cartoons. Normal subjects showed significantly greater comprehension and overt expression of affect than either group of retarded subjects. The retarded groups did not differ from each other along either dimension. Among all subjects, only the institutionalized retardates group did not show an increase in smiling and laughter with increased comprehension. It was concluded from this finding that the institutionalized retardate subjects tended to show more "inappropriate" affect. Among the institutionalized retardate subjects, length of institutionalization was significantly negatively related to humor comprehension, but was not related to appreciation. Within all three groups, comprehension was significantly positively related to mental age.

Dependent Variable in Research on Children's Humor

Along with the insufficient amount of attention devoted to the nature of the stimulus situation in humor research, many investigators have also failed to give serious consideration to the dependent variables used and to factors that might influence performance on those measures. One of the earliest methods used to study young children's humor was to observe spontaneous smiling and laughing in a natural environment. However, the value of the data obtained using this method clearly depends on the ability of the observer to determine the true laughter-evoking stimulus. While the observer may define the source of laughter in terms of his perception of the situation (and have perfect agreement with an independent observer), the child may have been responding to a different set of cues. This possibility is especially strong in a natural environment, where the total stimulus situation contributing to the humor experience is very complex. Even if we may assume perfect accuracy in the observer's identification of the stimulus elements to which the subject was responding, we have no more insight into what the child laughed at, or why, than is achieved upon the observation of laughter when he sees a cartoon or hears a joke. Thus, as with other approaches to studying children's humor, additional information must be obtained from the child himself about what he was laughing at and why he found it to be funny. However, since it is difficult to obtain this information from very young children, our ability to gain a complete understanding of the nature of the young child's humor response is more restricted. By using the observational technique as a preliminary step to suggest stimulus dimensions that might consequently be studied experimentally, this restriction might be partially overcome.

A similar set of problems confronts investigations using cartoons or jokes as humor stimuli. With respect to humor appreciation, the most important distinction to be drawn concerns the choice of an affective or intellectual index of appreciation. With the exception of Zigler et al. (1966a, 1967), who obtained both, investigators have relied on either overt expressions of affect (smiling and laughing) or some type of verbal rating to measure degree of perceived funniness. Each approach would appear to have its own set of methodological problems. Studies relying on overt expression of affect have assumed that laughter is higher on a continuum of intensity of affective expression than smiling, and therefore indicative of greater appreciation. While this assumption is probably a sound one in general, widespread individual differences exist in children's inclination toward overt affective expression of a humorous experience. While degree of comprehension undoubtedly plays some role in determining the degree of affective expression in all children, an affective rating scale probably describes the humor experience most satisfactorily for children showing the whole range of responsiveness from a blank expression to hearty laughter. For the consistently maximally or minimally expressive child, it may be that all of the humor stimuli are in fact perceived as very funny (or not funny), in which case the measuring instrument would be accurately reflecting the child's experience. A more likely possibility, however, is that the child does perceive some stimuli to be more humorous than others, but does not respond accordingly. This may be especially true for children showing consistent laughter in response to cartoons or jokes they do not understand. It may be that socially desirable responses are being evoked in such cases. That is, a child may laugh simply because he thinks that is what the experimenter expects or wants him to do. Moreover, since even the youngest school-age child has long known that cartoons and jokes are supposed to be funny and laughed at, the child may laugh simply because he has seen a cartoon or heard a joke.

Spontaneous expressions of affect are undoubtedly influenced by the mood or set the child brings to the experimental situation. Thus, if a young child is taken out of his classroom and placed in a room with a stranger, he may not feel enough at ease to be spontaneous and uninhibited in his emotional expression. A joke or cartoon that evokes only a smile in this situation might have produced great laughter in the child's classroom or at home. Similarly, anxiety might lead to the consistent "inappropriate" laughter noted above. In spite of our efforts to control or ease the child's anxiety, it is difficult to determine the extent to which anxiety may be boosting or toning down the child's general responsiveness.

The type of humor stimuli used may also play an important role in determining the adequacy of degree of affective expression as an index of appreciation. The writer has suggested elsewhere (McGhee, 1971a) that while different types of cartoons or jokes may be rated by a subject as being equally funny, some may be more conducive to laughter than others. Specifically, humor related to some salient emotional tendency or need (e.g., sexual or aggressive humor) may be more conducive to overt expressions of affect than more purely cognitive types of humor (e.g., incongruity humor) in which such themes are absent. The latter type of humor may be seen as being funny in the sense of being clever. Since the response here is more of an intellectual one than an affective one, an affective index of appreciation may not accurately reflect the child's reaction to the humor stimulus.

An approach using children's verbal ratings of funniness also encounters the problem of evoking socially desirable responses. However, this would appear to be reduced somewhat by the fact that the experimenter has given the subject a scale that allows a rating of "not funny at all." He may be less likely in this situation, then, to give it a high rating simply "because it's a joke" or because he thinks the experimenter wants him to rate it highly.

A problem characterizing the verbal rating approach, which is not shared by the affective approach, lies in the accurate usage of the rating scale provided. If, for example, a 5-point scale is utilized, the question arises as to the minimal age at which the child can accurately identify his feelings about a cartoon or joke and transfer them to a numerical scale. A younger child may find differentiation of the ends of the scale a relatively simple task, but have difficulty making the finer differentiations required by the scale points in between. Clearly, at some point we must use a dichotomous scale and simply ask the child whether the stimulus figure is funny or not.

It may be that each of these approaches is actually tapping very different aspects of the child's global humor response. Thus, affective measures may be greatly influenced by social variables, while intellectual ratings are influenced to some lesser extent. On the other hand, the latter measure may more accurately reflect the child's intellectual reaction than does the former. While the Zigler et al. (1966a) data on the relationship between these two measures is valuable, additional research efforts comparing the two approaches are needed. It may prove more fruitful to make predictions on the basis of a combined score than on either measure separately.

Measures of humor comprehension have been less consistent across different investigations. While most studies do not report the procedures used for scoring for comprehension, it appears that comprehension has typically been measured in terms of the degree of agreement of the subject's explanation with what the experimenter defines as the "point" of the joke or cartoon. This approach is weakened by the fact that criteria for different levels of comprehension must be established separately for each cartoon or joke. It would seem valuable to work toward the development of a general basis for scoring comprehension which might be applied to any set of humor stimuli. The studies of McGhee (1971a, 1971b) are based on such a general scoring system, although his system was designed for stimuli with a primarily cognitive (as defined earlier) basis for humor. A detailed description of this scoring system is presented in McGhee.[7]

UNRESOLVED THEORETICAL AND EMPIRICAL ISSUES

The most striking conclusion that may be drawn from this review is that theory has played a minimal role in generating those studies that have been completed, making it difficult to achieve any coherent integration of the data obtained. While classical psychoanalytic theory has generated a number of studies of humor in adults (see Gollob & Levine, 1967; Treadwell, 1967), only the Hetherington (1964) study has been completed with children. Arousal theories and classical gestalt theory also remain to be tested using children. Because of the prior inadequate attention given to theoretical issues, measurement problems, and stimulus, response, and subject variables in many previous studies, our understanding of children's humor has progressed very little. While the need for more theory-generated research is apparent, this may follow as a matter of course if more specifically developmental theories of humor are advanced.

Developmental theories of humor must be careful to distinguish between a theory of humor and a theory of laughter. While much of the data on very young children reviewed here is based on laughter, it is difficult to determine whether the laughter is indeed indicative of a humor experience. Since it seems clear that humor constitutes only one among several sources of laughter (see Flugel, 1954; Giles & Oxford, 1970), theorists must carefully define their criteria for a humor response. Specific issues that a developmental theory of humor must consider include (*a*) the onset of the capacity to experience humor, including organismic and environmental variables that influence its timing; (*b*) the nature of developmental changes in the humor response; (*c*) the ontogenetic relationship between smiling and laughter as a sign of pleasure in assimilation, body stimulation, etc., and an index of a humor experience; (*d*) the relative predominance of emotional or need-related versus more purely cognitive bases for humor, both initially and at subsequent developmental levels; (*e*) the role of cognitive, personality, and social variables in this developmental sequence; (*f*) the role of language acquisition in early humor; (*g*) the relationship between conceptual thinking and the capacity to experience humor; and (*h*) the role of intelligence and the acquisition of logical thinking in the child's ability to comprehend, appreciate, and generate humor.

---

[7] McGhee, P. E. Cognitive development and children's comprehension of humor. Unpublished doctoral dissertation, Ohio State University, 1968.

Before theories of humor may be satisfactorily tested, a number of basic methodological issues must be settled. These were discussed in the preceding section. Once we have achieved a better understanding of the independent and dependent variables used in research on humor, the study of children's humor might progress more rapidly upon the development of a standardized test (or tests), which might be available to all experimenters. This would maximize comparability of data and provide for a much greater integration of findings than is presently possible. Along with a sampling of varying content areas, types of humor, and modes of representation (cartoon, joke, etc.), this test should control for the level of cognitive demands made upon the subject for comprehension of the humor depicted, so that stimulus complexity might be adequately sampled.

The nature and role of arousal processes in humor remains untested. While the views of Berlyne (1960, 1969), Tomkins (1962), and Kagan (1967) have been discussed in this respect, there is clearly a need for a systematic investigation of the arousal continuum operative in various types of humor situations. In conjunction with this dimension, an investigation of the role of such factors as complexity, surprisingness, and degree of novelty or incongruity is needed. These factors, in combination with accompanying influences upon arousal, may play an important role in determining the effect of stimulus repetition on the child's humor response. While this dimension remains unexplored with children, degree of comprehension upon each encounter with the stimulus may also influence the effect of repetition upon appreciation.

Investigators have given little attention to the role of fantasy in the humor process. Many situations (e.g., a man being run over by a steam roller in a cartoon) appear to be humorous only because the child perceives cues that tell him that the event in question is only a play on reality; that is, exists only at a fantasy level. In such situations, removal of these cues may also remove the humor. Thus, it should be fruitful to study the role of fantasy in children's humor at different developmental levels.

The creative aspects of children's humor have also received little empirical attention. Most studies have simply recorded the subject's reactions to prearranged stimuli, and have not allowed the subject to generate humor on his own. Important areas for research along this dimension include (*a*) the age of onset of the capacity for self-generated humor; (*b*) the types of humor generated initially and at later developmental stages; (*c*) the role of cognitive development in initial and later phases of self-generated humor; (*d*) the relationship between social and personality variables and frequency of initiation of humor; and (*e*) the choice of measuring instruments for determining the capacity or tendency to create humor.

The influence of a social versus nonsocial environment upon children's humor also remains to be investigated. A valuable approach here might be to provide children with a wide array of potential laughter-evoking stimuli in a nonsocial setting (e.g., alone in a play room) and record the child's laughter or verbal rating in response to them. The influence of social facilitation upon measures of appreciation might then be determined by introducing the same set of stimuli under two sets of social conditions; namely, children present and either interacting or not interacting with the subject upon presentation of the stimuli. The need to maintain social desirability in test situations also constitutes an important potential social influence upon humor measures, and needs to be investigated.

Finally, additional areas needing further investigation include (*a*) the effect of anxiety on humor responsiveness; (*b*) sex differences; (*c*) the influence of the sex or behavior of the experimenter upon subjects' laughter or verbal ratings; (*d*) humor comprehension and appreciation in mentally retarded or emotionally disturbed children; (*e*) the role of personality and intellectual variables in individual differences in humor preferences; and (*f*) the relation between children's humor responsiveness and "healthy" personality development.

REFERENCES

Bateson, G. The position of humor in human communication. In J. Levine (Ed.), *Motivation in humor*. New York: Atherton, 1969.

Behar, I., & Bevan, W. The perception of incongruity by young children. *Acta Psychologica*, 1956, **12**, 342–348.

Berlyne, D. E. *Conflict, arousal and curiosity*. New York: McGraw-Hill, 1960.

Berlyne, D. E. Arousal and reinforcement. *Nebraska Symposium on Motivation*, 1967, **15**, 1–110.

Berlyne, D. E. Laughter, humor and play. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology*. Vol. 3. (2nd ed.) New York: Addison-Wesley, 1969.

Bever, T. Associations to stimulus-response theories of language. In T. R. Dixon & D. L. Horton (Eds.), *Verbal behavior and general behavior theory*. Englewood Cliffs, N. J.: Prentice-Hall, 1968.

Bird, G. E. An objective humor test for children. *Psychological Bulletin*, 1925, **22**, 137–138.

Brackett, C. W. Laughing and crying of preschool children. *Journal of Experimental Education*, 1933, **2**, 119–126.

Brackett, C. W. Laughing and crying of preschool children. *Monographs of the Society for Research in Child Development*, 1934, No. 14.

Brumbaugh, F. The place of humor in the curriculum. *Journal of Experimental Education*, 1940, **8**, 403–409.

Buhler, C. *The first year of life*. New York: John Day Company, 1930.

Ding, G. F., & Jersild, A. T. A study of the laughing and smiling of preschool children. *Journal of Genetic Psychology*, 1932, **40**, 452–472.

Eastman, M. *The sense of humor*. New York: Scribner, 1921.

Enders, A. C. A study of the laughter of the preschool child in the Merrill-Palmer nursery school. *Papers of the Michigan Academy of Science, Arts and Letters*, 1927, **8**, 341–356.

Flugel, J. C. Humor and laughter. In G. Lindzey (Ed.), *Handbook of social psychology*. Vol. 2. *Special fields and applications*. Reading, Mass.: Addison-Wesley, 1954.

Freud, S. *Wit and its relation to the unconscious*. New York: Moffat Ward, 1916.

Fry, W. F. *Sweet madness: A study of humor*. Palo Alto, Calif.: Pacific Books, 1963.

Gesell, A., & Ilg, F. *The child from five to ten*. New York: Harper and Brothers, 1946.

Giles, H., & Oxford, G. S. Towards a multidimensional theory of laughter causation and its social implications. *Bulletin of the British Psychological Society*, 1970, **23**, 97–105.

Gollob, H. F., & Levine, J. Distraction as a factor in the enjoyment of aggressive humor. *Journal of Personality and Social Psychology*, 1967, **5**, 368–372.

Graham, L. R. The maturational factor in humor. *Journal of Clinical Psychology*, 1958, **14**, 326–328.

Grotjahn, M. *Beyond laughter*. New York: McGraw-Hill, 1957.

Hall, G. S., & Allin, A. The psychology of tickling, laughing and the comic. *American Journal of Psychology*, 1897, **9**, 1–41.

Helmers, H. *Sprache und humor des kindes*. Stuttgart, Germany: Ernst Klett Verlag, 1965.

Hetherington, E. M. Humor preferences in normal and physically handicapped children. *Journal of Abnormal and Social Psychology*, 1964, **69**, 694–696.

Jones, M. C. The development of early behavior patterns in young children. *Pedagological Seminary*, 1926, **33**, 537–585.

Justin, F. A genetic study of laughter provoking stimuli. *Child Development*, 1932, **3**, 114–136.

Kagan, J. On the need for relativism. *American Psychologist*, 1967, **22**, 131–143.

Kenderdine, M. Laughter in the pre-school child. *Child Development*, 1931, **2**, 228–230.

Kimmins, C. W. *The springs of laughter*. London: Methuen, 1928.

Koestler, A. *The act of creation*. New York: Macmillan, 1964.

Kreitler, H., & Kreitler, S. Dependence of laughter on cognitive strategies. *Merrill-Palmer Quarterly*, 1970, **16**, 163–177.

Kris, E. Ego development and the comic. *International Journal of Psychoanalysis*, 1938, **19**, 77–90.

Laing, A. The sense of humor in childhood and adolescence. *British Journal of Educational Psychology*, 1939, **9**, 201.

Levine, J. Humor and play in sports. In R. Slovenko & J. A. Knight (Eds.), *Motivation and play, games and sports*. Springfield, Ill.: Charles C Thomas, 1967.

Maier, N. R. F. A Gestalt theory of humor. *British Journal of Psychology*, 1932, **23**, 69–74.

McGhee, P. E. Cognitive development and children's comprehension of humor. *Child Development*, 1971, **42**, 123–138. (a)

McGhee, P. E. The role of operational thinking in children's comprehension and appreciation of humor. *Child Development*, 1971, in press. (b)

McGhee, P. E. On the cognitive origins of incongruity humor: Fantasy-assimilation versus reality-assimilation. In J. H. Goldstein & P. E. McGhee (Eds.), *The psychology of humor*. New York: Academic Press, in press.

Piaget, J. *Psychology of intelligence*. London: Routledge and Kegan Paul, 1950.

Piddington, R. *The psychology of laughter: A study in social adaptation*. London: Figurehead, 1933.

Schaffer, L. F. *Children's interpretations of cartoons*. (Contributions to Education, No. 429) New York: Teacher's College, Columbia University, 1930.

STRICKLAND, J. F. The effect of motivational arousal on humor preferences. *Journal of Abnormal and Social Psychology*, 1959, **59**, 278–281.

TOMKINS, S. *Affect, imagery, consciousness*. Vol 1. *The positive affects*. New York: Springer, 1962.

TREADWELL, Y. Bibliography of empirical studies of wit and humor. *Psychological Reports*, 1967, **20**, 1079–1083.

VALENTINE, C. W. *The psychology of early childhood*. London: Methuen, 1942.

WALKER, M. A., & WASHBURN, M. F. The Healy-Fernald Picture Completion Test as a test of the perception of the comic. *American Journal of Psychology*, 1919, **30**, 304–307.

WASHBURN, R. W. A study of the smiling and laughing of infants in the first year of life. *Genetic Psychology Monographs*, 1929, **6**(5, 6), 397–535.

WELLS, R. E. A study of tastes in humorous literature among pupils of junior and senior high schools. *Journal of Educational Research*, 1934, **28**, 81–91.

WILLIAMS, J. M. An experimental and theoretical study of humor in children. *British Journal of Educational Psychology*, 1946, **16**, 43–44.

WOLFENSTEIN, M. Children's understanding of jokes. In, *Psychoanalytic study of the child*. Vol. 9. New York: International Universities Press, 1953.

WOLFENSTEIN, M. *Children humor*. Glencoe, Ill.: The Free Press, 1954.

ZIGLER, E., LEVINE, J., & GOULD, L. Cognitive processes in the development of children's appreciation of humor. *Child Development*, 1966, **37**, 507–518. (a)

ZIGLER, E., LEVINE, J., & GOULD, L. The humor response of normal, institutionalized retarded and noninstitutionalized retarded children. *American Journal of Mental Deficiency*, 1966, **71**, 472–480. (b)

ZIGLER, E., LEVINE, J., & GOULD, L. Cognitive challenge as a factor in children's humor appreciation. *Journal of Personality and Social Psychology*, 1967, **6**, 332–336.

(Received March 31, 1970)

Psychological Bulletin
1971, Vol. 76, No. 5, 349-364

THE BOGUS PIPELINE:
A NEW PARADIGM FOR MEASURING AFFECT AND ATTITUDE[1]

EDWARD E. JONES[2] AND HAROLD SIGALL
Duke University *University of Rochester*

If subjects can be convinced that a physiological monitoring device is able to measure both the amplitude and direction of emotional response, their subsequent attempts to predict what the machine says about their attitudes should be uncontaminated by many of the biases that obscure paper-and-pencil measures. This assumes that subjects do not want to be second-guessed by a machine. Four experiments are described in which variations of this "bogus pipeline" paradigm were employed. In each case it proved quite possible to validate the device in the subject's eyes and to extract theoretically meaningful estimation measures of interpersonal attraction. It is not clear whether the differences between such estimation measures and the standard verbal measures of attitude are based simply on a lie detection principle, on a principle of error preference, or on the fact that the affective component of a complex attitude has been made salient. For whatever reason or reasons, subjects attached to the bogus device appear much more ready to express negative affect in experimental settings where one might normally expect the inhibition of such feelings, for instance, toward Negroes and the physically handicapped.

Many psychologists for many years, beset with the vexing difficulties associated with inferring true feelings from behavior, must have had fantasies about discovering a direct pipeline to the soul (or some nearby location). Wouldn't it be nice if people really did wear their hearts on their sleeves? Or even if we had access to reliable and valid physiological indicators of attraction and dislike, approach and avoidance, or belief and disbelief? But alas (and some would say, thank goodness), no one has yet achieved this last breakthrough on the path to invasion of privacy. The present study describes a paradigm that appears to bypass some of the major problems of inference from behavior and poses new problems for substantive research along the way. The paradigm is based on the simple premise that no one wants to be second-guessed by a machine. If a person could be convinced that we *do* have a machine that precisely measures attitudinal direction and intensity, we assume that he would be motivated to predict accurately what the machine is saying about him. The bogus pipeline paradigm involves, then, the use of a device or machine that purportedly measures one's true feelings about a person or an issue, some means of validating the machine in the subject's eye, and finally the dependent variable itself: a prediction by the subject of the machine's telltale reading.

Before proceeding with a detailed discussion of this paradigm and its possibilities, it may be useful to place the procedure in the broader context of measurement in social psychology. Throughout much of the 1930s it could be reasonably argued that the measurement of attitudes was the central preoccupation of social psychologists. There was little theory to speak of, only an occasional experiment to report, and the notion of theoretically relevant situational manipulations was only beginning to take hold in the aftermath of Lewin, Lippitt, and White's (1939) studies on autocratic and democratic leadership. With the development of this manipulative experimental tradition during the 1940s and 1950s, the interest in attitude measurement was pushed from center stage. Especially the Festingerians (cf. McGuire, 1969) lavished great care on the independent variables being manipulated, while the dependent vari-

---

[1] The research reported in this paper was supported by National Science Foundation Grant G8857 to the first author and by National Institute of Mental Health Grant MH17180-01 to the second author.

[2] Requests for reprints should be sent to Edward E. Jones, Department of Psychology, Duke University, Durham, North Carolina 27706.

ables used were almost a casual afterthought. Instead of exploring the antecedent conditions that affected some standardized, reliable, and socially important dependent variable, experimenters in the Festinger tradition typically asked simple, direct, tailor-made questions to get at the effects of theoretically relevant situational variations. The consequences of elaborate experimental scenarios were often measured by a check on an *n*-point scale from "like" to "dislike" or a decision to wait alone versus wait with somebody else. Standardization, multimethod triangulation (cf. Campbell & Fiske, 1959), and cross-situational generality were typically sacrificed in favor of contextual relevance—measures introduced in as unobtrusive, "natural," and fitting a manner as possible.

Without in any way denying the importance of other research styles and strategies (e.g., McGuire's, 1969, Hovlanders with their search for interesting antecedents of standard dependent variables), we think the emphasis on contextual relevance has been generally well placed. It is hard to quarrel with the notion that the subject should respond to a reasonable question that fits into the cover story defining the purposes of the research, and that he not be dragged into a morass of complex and obtrusive psychometric measures. But there are obviously problems in assessing the answers that subjects give to any research question, and many of these problems are exacerbated by the one-shot direct probe. In the following section, we review some of these difficulties.

Potential Difficulties with Rating-Scale Measures of "True Feelings"

While many of the following difficulties apply to the general case of attitude measurement, we turn to the attraction area for our examples and consider ratings made by the subject of a target person.

1. *Generosity effect*. There is a common tendency for raters to give target persons the benefit of the doubt in recording judgments with obvious evaluative implications. On the kind of antonym scales often used in impression research, for example, it is rare to find subjects using the negative halves of scales like generous–stingy, wise–foolish, or friendly–hostile. The same is true, though to a smaller extent, with Likert-type items involving statements with which the subject is asked to express degree of agreement or disagreement (e.g., "She is rather on the stingy side"). This general tendency to give positive ratings is probably an expression of the realization that it never pays to risk giving offense when it is easy not to. We assume that many subjects do not want to appear odd or extreme and choose a moderately positive rating as a "place to hide" on the scale.

2. *Evaluation apprehension*. This latter consideration merges into a more general interest of the subject to be evaluated by the experimenter as normal, mature, perceptive, and full of integrity. "Evaluation apprehension" has been highlighted by Rosenberg (1965) and by Sigall, Aronson, and Van Hoose (1970) as a potential source of artifactual error in measuring the dependent variable in an experiment. Similarly, Edwards (1957) and Crowne and Marlowe (1964) have noted that the social desirability of a response may affect whether or not it is emitted. Depending on the subject's interpretation of the situation, his desires to be well evaluated may lead him to play it safe, to show how mature he is by not responding with a negative rating to a critical attack, or to show how perceptive he is by letting his ratings reflect small nuances in the target person's behavior. In short, the subject's evaluation apprehension may have diverse and context-specific effects on his ratings. In general, the subject will try to respond as he thinks a mature and rational person would.

3. *Experimenter demand*. Closely related to the notion of evaluation apprehension is the tendency that Orne (1962) and Riecken (1962) have posited for subjects to try to please the experimenter or make a special effort to confirm what they perceive to be his hypothesis. To the extent that this tendency is operative, it may obviously affect the subject's ratings of the target person—with or without his conscious realization.

4. *Thoughtlessness*. A major difficulty with paper-and-pencil measures of all sorts is that it is too easy not to care, not to be concerned about the validity of one's answers. There is no clear penalty for casual sloppiness, for

checking only the extremes or only the neutral points, rather than taking the trouble to be discerning.

5. *Errors of "psychologic."* One of the classically noted rating errors is that of the halo effect—the tendency for ratings on different dimensions to be more highly correlated than warranted by objective data. Halo effects are often seen in the evaluative domain, so that a target person who is rated positively on one trait is likely to be rated positively on other traits. But halo effects may also reflect more subtle rating biases that map the general connotative relations between trait names or the implicit personality theories held by the rater. Thus, if the rater believes that people who are generous also tend to be warm, he will rate the target person similarly on both dimensions—in spite of the fact that the target person may have behaved in a warm and stingy manner. A different kind of psychologic error may be seen in the rater's tendency to restore balance in his ratings, to rate friends of friends and enemies of enemies more positively than they deserve. Both kinds of error may be more common with paper-and-pencil rating measures than with measures that commit the subject to action or reflect an involving social relationship. Once again, it is easier on such measures, perhaps, to rate people the way they should ("logically") be rated rather than in line with one's true inner feelings.

Difficulties with Behavioral and "Behavioroid" Measures of Attraction

Strictly speaking, all measurement in psychology involves measurement of behavior, but one can conceive of a continuum of consequentiality for the subject (Aronson & Carlsmith, 1969). At one end of the continuum would be those actions that commit the subject to costly, effortful, long-range, or for any other reason personally important consequences. At the other end might be the anonymous questionnaire. Since truly consequential behavior measures are typically difficult and impractical to arrange in a psychology experiment, investigators have often resorted to what Aronson and Carlsmith (1969) have called "behavioroid" measures—distinguished from true behavioral measures in that the subject commits himself to consequential action without actually going through with it. From the subject's point of view, a behavioroid measure is the same as a behavioral measure at the time of commitment. If properly convinced, he truly expects to serve under the group member he votes for as leader (Schachter, 1951), or to escort visiting Negroes around campus (Marlowe, Frager, & Nuttall, 1965), or to have a large "drive carefully" sign displayed on his front lawn (Freedman & Fraser, 1966). The important consideration is that he expects to have to live with the very real consequences of his behavior, and therefore is hardly likely to take his response lightly or be much affected by evaluation apprehension or experimenter demand.

The most obvious advantage of behavioral and behavioroid measures of attraction is the likelihood that the subject will respond in a way more readily generalizable to the involving situations of real life. But there are disadvantages as well. If the measure is behavioroid, there is always the lingering suspicion that the subject does not really expect to follow through on his commitment. A more important problem is that behavior (and behavioral commitment) can be multiply or alternatively determined. In studies where the behavior itself is the sufficient focus of study (as in the effects of social pressure on volunteering behavior, or the role of frustration in determining the sociometric rejection of ethnic minorities), this problem is not of overriding importance. When behavior or effortful commitment is used as an index of attraction, however, the problem may be acute. A requirement of sociometric measures, according to Moreno (1934), is that subjects make their choices fully understanding that they can lead to restructuring the group. The original sociometric test was, then, a behavioroid if not a behavioral measure. But one can imagine many reasons that might underly a sociometric choice: The chooser feels that the group should be a mixture of good guys and bad guys, and since others are likely to choose nothing but good guys, it is up to him to choose a bad one; the chooser feels sorry for someone he expects others to reject (Bramel, 1969); the chooser may be more fearful of the target of his choice than of others and be trying to appease him.

Similar difficulties arise with other behavioral measures of attraction. We may reward another, not because we like him, but because we feel sorry for him. We may shock the student in a teacher–student paradigm, not because we dislike him, but because we think he is intelligent enough to profit by clear feedback. We may agree to circulate petitions or make phone calls for another person because it bolsters our own self-esteem to serve a worthy cause, rather than because we like the requester. We may prefer to meet again with someone to convert him to our position, rather than because we find him especially attractive. As the consequences of behavior become more important, the person will presumably pay less and less attention to the picture he presents to the experimenter. Nevertheless, as Cook and Selltiz (1964) pointed out, even in a genuine behavioral setting in which choices or commitments are expected to have consequences, "an individual may be motivated to give a response that differs from his spontaneous private one, in order to present himself . . . as unprejudiced or to maintain his own image of himself as one who behaves in an unprejudiced way [p. 45]."

The division of measures into paper-and-pencil and behavioral hardly exhausts the various ways in which measures of affect and attitude can be measured. Both behavioral and questionnaire measures may be direct or indirect. A direct measure (to paraphrase Scott, 1969) is one in which the subject and the investigator are assumed to attach the same meaning to the question or stimulus and infer the same significance of the answer or response. There are many ways to disguise a measure of attitude or attraction. Some of these indirect measuring techniques are administered as part of a mild diversion, and some are projective and involve displacements of stimuli or responses. As an example of the former, students of interpersonal attraction often stress their interest in how first impressions are formed, but develop indexes of attraction from the first-impression ratings by adding scores on all the evaluative traits (e.g., Jones & deCharms, 1957; Jones, Stires, Shaver, & Harris, 1968). This is not too different from the common procedure of measuring attitudes toward an object by measuring beliefs about it. Also in the attitude realm, similar but more clearly projective use may be made of what Campbell (1950) has called "disguised-structured" tests. An example here would be the test that purports to measure knowledge or information about a domain of attitudinally relevant facts, which is scored in terms of the prevalent direction of bias in the wrong answers given.

As an example of projective measures of attraction, Gerard (1965) used a Group Picture Impressions Test devised by Libo [3] to obtain an indirect measure of subjects' evaluations of other subjects in a conformity situation. The test involves writing stories to four stylized pictures showing an individual in different group situations. These stories were then content-analyzed to determine approach or avoidance reactions to others by the main character in each of the pictures. The resulting algebraic sum of approach minus avoidance statements was used as a measure of each subject's attraction to those in his present situation, with results confirming Gerard's particular theoretical hypothesis.

The problem with indirect or projective measures of attraction is that they require careful validation, and the experimental situations in which they are eventually used may differ in important ways from the original validation settings. Even with tests of fairly high validity there is bound to be substantial error variance contributed by idiosyncratic interpretations of the items or pictures, by level of verbalization, and by various response styles.

Physiological Measures of Attitude and Attraction

It is hardly surprising that for many of the reasons implied by the above discussion, investigators have sought to discover physiological indicators of attitude. It is usually assumed that such indicators are less subject than verbal responses to voluntary distortions and defensive denial. Although they were well aware of previous unsuccessful attempts to use the galvanic skin response (GSR) in the diagnosis of social attitudes, Rankin and Campbell (1955) obtained highly significant GSR-level differences between subjects responding to contact with white versus Negro experimenters. Although Porier and Lott (1967) were unable to replicate this finding in a better controlled study, they did find a low but significant correlation between GSR reactivity to Negro contact and ethnocentrism scale score. Westie and DeFleur (1959) recorded finger-pulse volume as well as GSR as subjects were shown pictures of Negroes and whites flashed on a screen. Prejudiced subjects (as measured by a social distance questionnaire) showed greater GSR reactions but lowered pulse rates to slides depicting Negroes. In a similar study, Cooper (1959) found that GSR magnitude to ethnic names like "Mexicans," "Swedes," and "Japanese" could be predicted from verbal reactions to the same labels. Hess (1965) reported that pupillary dilation seems to correlate with positive affect.

---

[3] Libo, D. Measuring group cohesiveness. Ann Arbor, Mich.: Research Center for Group Dynamics, 1953.

These and similar studies (cf. Leiderman & Shapiro, 1964; Shapiro & Crider, 1969) exploring autonomic indicators of affective arousal may shed some light on the structure and patterning of attitudes when the indicators are viewed in the context of specific experimental procedures and related to the pattern of conventional verbal responses to the same situation. As a direct measure of "true" attitude, however, physiological indicators are judged to be crude and responsive to too many variables. For one thing, there is increasing evidence that physiological variables are affected by such cognitive factors as instructional set, suggestion, and cognitive dissonance. Gerard (1964) measured finger-pulse amplitude changes during a period in which subjects were to decide which of two paintings they wished to have. By assuming that pulse rate should be high when the subject was experiencing dissonance, Gerard was able to provide a rough physiological charting of the decision process featuring dissonance elevation and reduction. Zimbardo, Cohen, Weisenberg, Dworkin, and Firestone (1969) also showed that predicted dissonance arousal and reduction effects can be tracked by changes in GSR. In one sense these studies emphasize the considerable promise of using physiological indicators to reflect cognitive changes, but they also suggest that it is misleading to contrast voluntary control over verbal responses with involuntary physiological indicators.

A second and more obvious problem with the use of physiological measures of attitude and attraction is the difficulty of inferring affective *content* or *direction* from such measures. As Westie and DeFleur (1959) noted:

> It cannot be assumed that conspicuous physiological changes manifested by an S [subject] upon being exposed to a stimulus object indicates simple favorability or unfavorability toward the attitude object, regardless of his scale-measured attitudes toward that object. It is easy to fall into thinking, especially when the stimuli are "race objects," that high response indicates a negative attitude [p. 346].

Shapiro and Crider (1969) raised a number of doubts concerning the use of physiological measures even in the detection of stress, arousal, tension, or activation. It is obviously hazardous in the extreme to make the still further assumptions that would be required to use GSR or pulse-rate measures as indicators of the direction and intensity of attitude.

For purposes of the bogus pipeline measure now to be described, however, the crucial consideration is the *plausibility* of having physiological measures of one's true attitudes and feelings. Here our task is greatly facilitated by public knowledge and stereotypes about the "lie detector" and its use in criminal investigations. According to Sternbach, Gustafson, and Colier (1962), lie detection by polygraph is still in the state of very primitive art, suffused by unfounded and often fraudulent claims concerning reliability and validity. But Lee (1953) emphasized the preventive utility of lie detection devices and operators because of the fact that many would-be larcenists and perjurers believe in the mystique and infallibility of lie detection. He retold the story of the prince who rounded up suspects and told them that each was to go into the next room blindfolded and pull the tail of the ass therein. The ass would bray when his tail was pulled by the guilty party. The prince examined the hands of each suspect after they had completed, in turn, their quiet visit to the next room. The one suspect whose hands were not black was accused of being the guilty party. The ass's tail had been dusted with black powder, and the prince assumed that the guilty party would only pretend to pull the telltale tail. The moral of this story for our purposes is that if people can be made to believe that there are devices reflecting their true inner attitudes or feelings, the measurement possibilities are almost limitless.

VERSIONS OF THE BOGUS PIPELINE

The idea of exploiting false physiological feedback in a social psychology experiment is not new. Unlike the present proposed use of the bogus pipeline concept, however, earlier uses have manipulated feedback as an independent experimental variable. Perhaps Gerard[4] was the first to see the possibilities of such feedback in controlling certain assumptions that the subject was making about himself and his proclivities. Gerald attached electrodes to the forearms of subjects in a standard conformity setting requiring judgments of the larger of two multipointed stars. Each subject was led to believe that the electrodes were attached to highly sensitive electromyographic equipment that was capable of measuring "implicit muscle movements" in the forearm. Since the response task was to depress a button with either the left or the right index finger, this measure could plausibly reflect the subject's first impulse to choose one or the other star in the situation. By letting the subject see the first impulses of others as well as his own first impulse (all fictitious, of course) on meters in the subject's display panel, it was possible to convince the subject that he had a tendency to yield or be autonomous, regardless of his overt, conscious preferences in the situation.

Bramel (1962) convinced subjects that degree of homosexual arousal to pictures of men in various states of undress was being reflected on the meter in front of them. Brehm and Behar (1966) used a similar feedback technique to convince their subjects that they were or were not sexually aroused to line drawings of a turtle, cup, flag, table, and rosebush. Valins (1966) played back amplified (fictitious) heart beats to subjects while they observed slides depicting nude women. Gerard and Rabbie (1961) and Behar (1967) were successful in making subjects think that they were more or less fearful in a shock-anticipation setting where they were allowed to observe their own (bogus) GSR readings. In each of these studies, the feedback information was a major part of the independent variable manipulation defining the various experimental conditions. The feedback was embedded in a reasonable cover story, and there was no apparent problem of suspicion.

We consider the bogus pipeline to be a dependent variable paradigm, rather than a specific device or machine for measuring attitudes and feelings. As indicated in the introduction, the paradigm involves convincing the subject that there is a machine that provides valid indication of his true feelings, hooking him up to such a machine, introducing theoretically interesting stimulus conditions to the subject while he is hooked up, and finally asking the subject to estimate the machine reading in terms of a scale presumably paralleling the summary readout of the machine. Thus far, we have conducted experiments with two different kinds of feedback machinery that illustrate the pipeline paradigm: the directional or "differentiated" GSR and the electromyograph. We present some procedural details and results for each of these devices to demonstrate their utility and certain special problems associated with their use.

Galvanic Skin Response "Breakthrough"

Jones, Bell, and Aronson (1971) conducted an experiment investigating the combined effects of opinion similarity and being liked on the reciprocation of attraction. In short, how do we feel about similar people who like versus dislike us? How about dissimilar likers and dislikers? Jones et al. predicted that subjects will be (*a*) more attracted to a dissimilar liker than a similar liker and (*b*) less attracted to a similar disliker than a dissimilar disliker. Their reason for predicting such a main effect for dissimilarity was the special significance attached to expectations that are violated in a pleasant (being liked by a dissimilar person) versus an unpleasant (being disliked by a similar person) direction. We should be especially inclined to reciprocate attraction when we have unexpectedly gained the admiration of a dissimilar other, and we should be especially prone to dislike someone who dislikes us in spite of the opinions we share in common. When no information about being liked or disliked is provided, we should simply like the similar better than the dissimilar other.

---

[4] Gerard, H. B. Acts, attitudes and conformity. (Symposia Study Series No. 4) Washington, D. C.: National Institute of Social and Behavioral Science, 1960.

The experiment was presented to the subjects as a study of the process of impression formation. Through a series of complicated but plausible procedures, each subject learned whether another subject was similar or dissimilar in attitudes and then learned whether the other subject liked or disliked her after watching her being interviewed by the experimenter. In some cases (control condition), information about being liked was not provided. As a final step in the experiment, the subject observed a videotape of the other subject being interviewed. The interview had been previously taped by an experimental confederate. At no point during the experiment was the confederate actually present.

Throughout her observation of this last videotape, the subject was hooked up to a GSR machine, the properties of which had been explained to her previously. The main dependent variable measure of attraction was obtained by asking the subject to try to predict the final meter reading of the machine as she thought about the other subject. By this time, of course, she knew whether the other subject was similar or not and whether the other subject liked her after observing the interview. The machine was originally presented to the subjects as a "new kind of galvanic skin response—you know, GSR—machine." The novel feature of the machine was allegedly the result of a "breakthrough," enabling it to measure not only the magnitude of emotional arousal but to distinguish between positive and negative emotional states, between like and dislike. The subject was then shown a standard GSR machine, and her attention was drawn to the summarizing dial that reflected an "integration" of the subject's feelings at a given point in time on a scale ranging from −3000 to 3000.

TABLE 1

ESTIMATES OF "OWN GSR" ATTRACTION READING (Experiment 1)

Attitude	Being liked			
	Yes	No	Uncertain	Total
Similar	1396	−642	1392	715
Dissimilar	1852	− 38	981	932
Total	1624	−340	1187	

Note.—The meter dial on which estimates were to be based ran from −3000 to 3000.

A number of steps were taken to validate the meter in the subject's eyes. She observed the accomplice (again via prerecorded tape) apparently hooked up to the GSR while thinking about her "best friend" and then while considering someone she disliked or was mad at, during which time there were appropriate movements of the visible meter toward the positive and then toward the negative end of the dial. She herself went through the same best friend, disliked-person procedure and, though not allowed to see her own meter movements, the experimenter assured her that the machine appeared to be in working order. Finally, in order to encourage subjects to do their best in estimating the meter readings, the experimenter stressed that the accomplice had been very accurate in estimating *her* own meter readings, and the experimenter was convinced and hopeful that the subject could be accurate too.

Although the meter estimation was the primary measure of the subject's attraction for the accomplice, each subject was also asked to fill out a first impression questionnaire, the major portion of which consisted of twelve 15-point trait rating scales, each expressed in antonym form (intelligent–unintelligent, friendly–unfriendly, etc.)

The results of this experiment are discussed in detail in the original publication, and for present purposes their methodological implications are stressed. The mean estimates of "own" GSR attraction reading are presented in Table 1. Although the predicted interaction is only of borderline significance ($p < .07$), the means are all in the proper order: If no information about being liked is reported, the similar accomplice is better liked than the dissimilar one; if information about being liked *or* disliked is provided, the dissimilar person is preferred. Although in all cases the questionnaire followed the GSR estimate, and might therefore be expected to be somewhat contaminated by it, the impression ratings did not confirm the prediction or even fall in the right direction. One cluster of items designed to measure "social attraction" showed a general overall (nonsignificant) preference for the similar accomplice. With both measures, not surprisingly, there was a highly signficant preference for the liking over the disliking accomplice.

In a replication of this experiment (with a different accomplice made up to be less attractive), this very strong preference for the liking accomplice was maintained again for both measures. However, the effects of similarity–dissimilarity in the first experiment did not replicate. Once again, the meter-estimation measure and the social attraction cluster ratings yielded different results. This time, however, the subjects gave the similar accomplice much higher social attraction ratings than the dissimilar accomplice—regardless of her liking or disliking the subject. The meter-estimation data showed no effects of similarity whatsoever. In spite of this failure to replicate, in both experiments the meter-estimation data were more clearly in the direction of the hypothesis than were the conventional questionnaire data. As Jones et al. (1971) argued in their discussion of the entire pattern of results, the meter-estimation measure performs more like a measure of how the subject really feels about the accomplice, whereas the questionnaire data seem to be more a reflection of how the subject thinks she *should* feel in such a situation.

Such a conclusion is consistent with, but hardly forced by, the data of their two experiments. At a more conservative and more descriptive level, there is evidence that the meter-estimation measure and the attraction ratings are differentially responsive to experimental conditions, but have some proportion of variance in common. In the first experiment, for example, the within-cell correlations between the meter estimations and the attraction ratings averaged .66. In the second experiment, the average intercorrelation was .81. Thus it is clear that attractiveness ratings are moderately congruent with estimated affective reaction, though these correlations may indeed be inflated by the aforementioned confounding of measure and sequence of administration. On the other hand, the meter-estimation averages fall into a different pattern than the attraction ratings, as we have noted, and this suggests that the meter-estimation technique, while more cumbersome and deceptive than the questionnaire method, is tapping a different aspect (a truer aspect?) of attraction. This suggestion is strongly supported by the data from an experiment by Cooper (1971), to which we now turn.

Attraction as a Product of Dissonance Reduction

Cooper was interested in testing the theoretical hypothesis that for postdecisional dissonance to occur, any unfavorable consequences must have been foreseeable at the time of the decision. Cooper chose to test this hypothesis using female subjects in an alleged study of the effects of cooperation and competition on interpersonal attraction. He designed a two-person cooperative game in which each partner was to take turns attempting to answer problems from an aptitude test. Each player was also to announce whether she thought she was probably correct or incorrect in her answer to each problem. An incentive system was explained that provided maximum credit for answers that were correct and claimed to be correct by the respondent. The system, in effect, penalized the team if either partner was overly timid (failing to recognize her correct answers) or excessively confident (claiming incorrect answers to be correct). Some subjects were then led to expect a timid partner, some an overconfident one. Experimental subjects were given the option of exchanging the assigned partner for another randomly chosen one (none did so); control subjects were given no such choice. In the play of the game, subjects learned, through controlled feedback, that their partner was in fact very timid, moderately timid, moderately overconfident, or very overconfident. After playing the game, each subject was provided the main dependent variable measure of attraction for her partner.

This main measure was again in the form of estimated readings of a GSR machine featuring directional as well as intensity information. Subjects appeared at each experimental session in groups of four to six, having volunteered for an experiment on "cooperative problem solving." It was explained that a new device would be used to measure how each person felt about her partner during the problem-solving game. The device was described as the differentiated GSR, whose properties were essentially the same as those attributed to the machine by Jones et al. (1971). As part of the overall strategy of validating the differentiated GSR in the subjects' eyes, they were ushered into a small adjacent control room and shown an imposing array of electrical and electronic equipment (actually consisting of the laboratory intercom patch board and a bank of preamplifiers). After they had returned to the experimental room, each subject was assigned to a desk, separated from her neighbor by a partition. On each desk was an electrode and earphones. The wire from the electrode ran through a wall conduit into the adjacent equipment room. When the electrodes were taped to the subject's hand, she would presumably be attached to the elaborate differentiated GSR equipment supposedly designed to provide simultaneous feedback measuring the affective responses of each subject.

The experimenter then asked each subject to listen to all further instructions through the earphones in front of her. Under the pretext of making minor adjustments in the differentiated GSR base line for individual subjects, the experimenter asked each subject in turn to consider her feelings about the statement, "Since it is known to be nonaddictive, the use of marijuana should be legalized." This statement had appeared as an item on an opinion questionnaire adminstered to all subjects earlier in the semester. With the results of each subject's answers to this questionnaire before him, the experimenter announced over the earphones that the first subject's meter reading was so and so, on a scale from zero (meaning extreme negative affect or disagreement) to 300 (which indicated extreme positive affect). He then asked the subject, "Does this sound accurate to you?" In the vast majority of cases, the subject said yes and often expressed amazement at the accuracy of the meter. An occasional subject demurred and suggested that the meter needed a little adjustment. In such cases the experimenter purported to adjust the base line and proceeded to a second opinion item, again announcing the machine's reading and asking the subject if this was accurate. In all cases this was sufficient. The experimenter repeated this procedure with the second, third, fourth, etc., subject. It should be noted that this procedure uses each subject's acquiescent response as evidence to validate the discernment of the machine in the eyes of the others. The effect was very powerful, and the subjects were characteristically quite surprised to learn after the experiment that the machine's properties were fictitious.

After these validation procedures, the experimenter gave subjects some crucial information about their partners to create the requisite expectations of timidity and overconfidence, and the problem-solving game was then played. Each subject thought she was "Subject C," and the experimenter conveyed the same pattern of self- and partner feedback in each session. By indicating whether or not the partner gave the correct answer, and whether or not she certified the answer as probably correct, the experimenter could create the experience of playing with a partner who was either on the timid or the overconfident side.

At the end of 15 problems, subjects were asked to estimate and record "what you feel the differentiated GSR meter should read as you think of your partner." The subject's response at this point was the main dependent variable measure of the experiment. Subsequent to this, subjects were also asked a series of questions about their partners' intelligence, whether they would like to work with them again, etc.

For a detailed presentation of the results, the reader is referred to the original article. It is sufficient for our present purposes to indicate that the theoretical predictions were confirmed in detail with the differentiated GSR measure. In the choice conditions (where the subject could have asked for another partner), (*a*) when timidity was expected, the extremely timid partner was liked better than a moderately timid partner; and (*b*) when overconfidence was expected, the extremely overconfident partner was liked better than the moderately overconfident one. In short, when the subject is personally responsible for choosing to play with a partner possessing known liabilities, she likes the partner more when these liabilities turn out to be fairly extreme. This is in spite of the fact, of course, that the extremely timid or the extremely overconfident partner causes the team to lose more money than their moderate counterparts. This pattern of results was reversed when the subject did not have any choice in the matter, or when the subject expected a timid partner and got an overconfident one, or expected an overconfident partner and got a timid one.

These results again provide a prima facie case for the utility of the bogus pipeline as a measure of attraction. Just as surely, however, they do not provide the crucial comparative information that such measures are better than a direct attraction question. Since Cooper was primarily interested in the theoretical question, the role of personal responsibility in dissonance reduction, he did not run control subjects whose feelings were assessed by standard rating techniques. The data he obtained from the subsequently administered questionnaire provided several indicators that would normally go along with an impression of attractiveness (e.g., "enjoy working with"). These verbal questionnaire responses did *not* confirm the dissonance prediction. Even in the choice conditions, subjects did not express any particular desire to work further with the extremely timid or overconfident partner, though the differentiated GSR showed their attraction to such a partner when her proclivities were expected. Since the verbal measures followed the meter-estimation measure, however, we cannot be sure that the former were not in some way contaminated by the latter. A more direct test of the difference between the bogus pipeline and standard rating procedures was of major interest in two studies by Sigall and Page, now to be reported.

Bogus Electromyograph and Reactions to Stigma

Sigall and Page[5] were interested in directly comparing the bogus pipeline with a standard rating procedure. To this end they selected a situation in which subjects might be expected to distort their responses under normal rating conditions. As noted earlier, subjects are reluctant to be negatively evaluative. In addition, Goffman (1963) has nicely documented the awkward and affected nature of interaction with stigmatized others. Sigall and Page confronted female subjects with an obnoxious male stimulus person who was presented as either stigmatized or normal, and assessed subjects' reactions to the stimulus person under bogus pipeline or rating conditions.

The procedure went as follows: Each female subject, and male accomplice posing as a subject, reported to a waiting room in the psychology laboratory. The accomplice either wore massive leg braces or appeared to be physically normal. A posted sign directed subjects not to talk with one another. Moments later a telephone in the waiting room began to ring. When it became apparent that no one would answer it, the confederate picked up the phone and was extremely rude and gratuitously abusive to the caller, who evidently had dialed a wrong number. Shortly thereafter the experimenter appeared, seemingly unaware of the phone call. He ushered the subject and confederate into a small experimental room and asked them to respond to a short opinion questionnaire that consisted of five innocuous items concerning movie and music preferences. The experimenter then left the room, and subjects indicated agreement or disagreement with the items on a 7-point scale. While the subject answered her questionnaire, the confederate acted as if he were doing the same; actually he was copying the subject's responses. When the experimenter returned he "randomly" selected the subject to accompany him into a second experimental room. She was instructed to bring her questionnaire with her.

The subject was seated before a console

---

[5] Sigall, H., and Page, R. Attenuation of the expression of affect. Manuscript in preparation.

A steering wheel extending from the console turned a pointer along a 7-point scale ranging from −3 to 3. The drawn scale was mounted on the console, and there was a slot at each end of the scale in which a label could be placed. A small metal box sat next to the console. This box, labeled "EMG," contained a meter with a 7-point scale identical to the one on the console. The meter itself was marked "EMG Output." Two skin electrodes were connected to the box by cable, and cables also ran from the box to impressive-looking electrical junk.

The experimenter described the apparatus as an adapted electromyograph. It was explained than an electromyograph (EMG) was a device of long standing, used to measure implicit muscle potentials. The experimenter noted that standard questionnaires used in psychological research involved a variety of difficulties and that the newly developed adapted EMG was an important breakthrough that circumvented many of the problems. He explained that when the subject held the steering wheel, and electrodes were attached to her forearms, the EMG would pick up her tendency to turn the wheel to any point on the scale. She was told that she actually did not have to turn the wheel, that the EMG screened out *gross* muscle movements and recorded the "first, undistorted" reaction. The electrical junk was described as a small computer, which integrated the electrophysiological input, and the EMG output meter would reflect a close approximation of that integration. Similarities to a lie detector were noted, and it was pointed out that the EMG was an improvement over the lie detector in that it was sensitive to direction as well as intensity of responses.

The experimenter then told the subject that occasionally base-line responses differed among individuals, in which case an adjustment of the EMG would be required. To check on the subject's base line, they would go over the items on the questionnaire and look at the correspondence between it and the EMG readings. The wheel was locked in place, "agree" and "disagree" labels were inserted in the console, and the electrodes were attached to the subject. She was to listen to the statements as they were read, to attend to the drawn scale, and to look at the EMG output readings if she wished. The experimenter read the first item and flipped a switch on the computer. The computer's lights flashed, and a mechanical noise was emitted. After a brief period the noise ceased, and the needle on the output meter swung to a number. Although these events appeared automatic to the subject, they were manipulated by the accomplice in the adjacent room. The experimenter continued by presenting the remaining items in similar fashion. On one item the experimenter suggested that the subject could, if she wished, try to fool the EMG either by exercising gross muscle movements or consciously concentrating in the direction opposite to her true position. The subject was reminded that since the EMG recorded implicit muscle movements, a reflection of her first, undistorted reaction, such strategy should not be successful. After all items had been completed, the subject was asked to produce the questionnaire, and the two sets of responses were compared. Of course, the EMG had been accurate in every case.

The experimenter then turned to the "purpose of the present experiment." He said that the study was concerned with social perception and that more information would be disclosed later. At this point he needed to get some information from the subject regarding her current perceptions about herself, the experimental situation, and the other subject. He indicated that while he would obtain this information directly from the EMG, he was also interested in discovering how sensitive people were to their own feelings—"whether they were in touch with how they felt." Thus, he introduced a shield to obstruct the subject's view of the meter, and told her to predict what she thought the meter would read. It was explained that predictive accuracy would be maximized if she relied on her undistorted reactions. He also told her that she could see how well she had predicted after all the items had been administered. The experimenter presented the items, recording the subject's prediction after each one, and pretending to note the EMG reading as well. The crucial items asked for the subject's evaluations of the accomplice. The subject indicated her personal liking for the confederate, and her agreement or disagreement as to the applicability of various traits, such as "obnoxious," "pleasant," or "pushy" to him.

In the rating conditions the computer and EMG were in view, but turned off. The wheel was free to be turned, and the electrodes were hidden. The subject responded by turning the wheel and orally stating the number to which the pointer had been moved. It was explained that concentration and attention remained higher when a physical task was involved than when a paper-and-pencil questionnaire was used. The remaining apparatus was described as a small computer used for data analysis. As in the EMG conditions, subjects were asked to present first reactions, not to distort responses, and to be honest.

A detailed accounting of results is in preparation (see Footnote 5). Here we simply note some of the findings. All subjects were convinced that the EMG did in fact possess the powers attributed to it. As expected, subjects in the EMG conditions tended to be unfavorable in their evaluations of the confederate, while those in the rating conditions tended to be favorable. This was manifested in relative terms and also reflected by the fact that on a scale which ranged from 3 to −3, mean responses in the rating condition were different in algebraic sign from mean responses in the EMG condition. The EMG subjects said they did not like the stimulus person; rating subjects said they did. He was evaluated as pushy, obnoxious, and insensitive in the EMG condition, but not in the rating condition. In sum, EMG subjects seem to have responded reasonably to the staged obnoxiousness of the accomplice, whereas subjects in the rating conditions appear to have been overly generous.

Sigall and Page also expected a greater discrepancy in evaluation between measurement conditions when the confederate was stigmatized. They felt that most subjects would be hesitant to derogate someone with a stigma and that therefore especially high evaluations would result in the stigma-rating condition. Subjects in this condition did provide the most positive responses, and, although the form of the interaction was realized, it failed to reach statistical significance. This leaves open the possibility that the EMG simply elicits more negative responses than usual rating procedures. The final experiment reported here provides some information on the viability of such an alternative.

New Look at Stereotypes

One of the findings recently reported by Karlins, Coffman, and Walters (1969) in their study of social stereotypes was that white, American subjects provided less favorable stereotypes of "Americans" and more favorable stereotypes of "Negroes" than was the case in earlier research. Without denying that those data may reflect a real change in attitudes over time, it is quite possible that such findings may be partly influenced by social desirability. Self-criticism by Americans and a tendency to express positive sentiments about Negroes seem to be active phenomena on many campuses. Such a climate easily could contribute to response distortion in a study of stereotypes.

A second experiment by Sigall and Page (1971) attempted to examine possible distortion by using the EMG. A 2 × 2 factorial design was employed in which half of 60 white male subjects indicated how characteristic each of a series of adjective traits was of "Americans"; the remaining subjects did the same in response to "Negroes." This variable was crosscut by assessment technique: As in the previous study, half the subjects responded by predicting their EMG readings, whereas the other half filled out standard rating forms.

The procedure used to "validate" the EMG was essentially identical to the one used in the experiment on stigma. However, since there was no need to expose the subject to a confederate, one variation was introduced. When the subject arrived, he filled out the small questionnaire. He was then escorted to the experimental room and seated facing the console, with his back to the door of the room. The completed questionnaire was placed on a table near the door, which was left slightly ajar so that an accomplice in the corridor secretly could copy the subject's answers. The accomplice then went to the control room, able to manipulate the EMG readings during validation. As far as the subject could tell, no one had the opportunity to see his questionnaire.

After the preliminaries (including validation in the EMG conditions), subjects were asked to think about "a group," either "Americans" or "Negroes." The labels "characteristic (3) and "uncharacteristic" (-3) were inserted in the slots on the console, and the experimenter proceeded to read 22 traits, one at a time.

A detailed report is presented in Sigall and Page (1971). Here we summarize some of the more striking results. Some traits were unaffected by measurement technique and yielded only a Negro-American main effect. For example, Americans were described as more practical and more materialistic, while Negroes were described as more musical.

More interesting, especially for methodological considerations, were those traits that resulted in interactions. These traits tended to be highly affect-laden, and the form of the interaction was consistent over traits: "Americans" were described more favorably in the EMG than in the rating condition; "Negroes" were described more favorably in the rating than in the EMG condition. Two examples serve to illustrate. On a 7-point scale that extended from -3 to 3, "honest" was reported as characteristic (.60) of Americans in the EMG condition but as uncharacteristic ($-.27$) of Americans in the rating condition, while the same trait was viewed as uncharacteristic ($-.33$) of Negroes in the EMG condition but characteristic (.67) in the rating condition. This interaction was highly significant ($p < .001$). "Lazy" was reported as uncharacteristic ($-.80$) of Americans in the EMG condition and uncharacteristic ($-.60$), but less so, in the rating condition. "Lazy" was uncharacteristic ($-.73$) of Negroes in the rating condition, but characteristic (.60) in the EMG condition.

Regarding the question of whether the EMG simply elicits more negative responses, it should be noted that one of the groups, Americans, received more favorable evaluations in the EMG than in the rating condition.

Discussion and Conclusions

The examples reviewed in this report suggest a few of the various uses to which the bogus pipeline paradigm can be put. We wish to emphasize the rather unrestricted adaptability of the paradigm and to reiterate that particular procedures and rationales employed will vary with the investigator's interests and resources. As we penetrate more and more deeply into the layman's assumptions about physiological processes, it may be important to distinguish between bogus autonomic indicators and those alleged to reflect implicit muscle movements. For the moment, we are inclined to treat these as sufficiently equivalent to lump them together as one general paradigm.

While the above examples are encouraging and suggestive, they do not provide very precise information concerning the particular mensurational advantages of using bogus physiological indicators. This is especially true of the first two studies where there is no direct comparative information on the relation between meter estimates and straightforward verbal measures of attraction. The meter-estimate measure seems to give results that are more in line than the questionnaire measure with theoretical predictions of variations in underlying attraction, but there is no way of knowing whether this is because the meter-estimation measure is more penetrating or because it was administered first and therefore assigned greater importance by the subjects. The two Sigall and Page studies do provide direct comparative information, but it is far from entirely clear just what is being compared.

One of the strongest reasons for preferring a bogus physiological measure to a questionnaire rating is precisely that the rating is potentially vulnerable to so many extraneous and artifactual influences. It seems probable that different wordings and contexts of administration would produce rating results differing from each other and showing greater or lesser differences from meter-estimation measures. We know from the Sigall studies that electromyographic reading estimates differ systematically and understandably from paper-and-pencil ratings made on a scale with the same range and verbal label. But we do not know if the observed differences would remain or be as strong if, instead of using a standard paper-and-pencil rating as our comparison standard, Sigall had taken his subjects aside and said, "Look I want you to be completely candid and tell me what you really think about that guy," or "Give me a rating that reflects your gut reaction" (cf. Aronson & Linder, 1965).

This discussion moves us toward the more fundamental question about *why* the observed differences between verbal ratings and bogus physiological measures occur. Here we face a number of reasonable alternatives. The first and most straightforward of these is that bogus pipeline measures *operate like lie detection devices in facilitating scrupulously truthful reporting*. To embrace this alternative is to accept the assumption that subjects do not typically tell us the truth when they report their attitudes and feelings about sensitive matters, or those where private feelings might be expected to differ from a socially respected norm. This is not an altogether unreasonable assumption—especially if we remember that all it takes is a few subjects who conceal the extremity of their feelings in the rating condition to produce a reliable difference between condition means.

A variant of this alternative is to posit a *preference for negative over positive errors*. Let us assume that subjects have no univocal or precise response when asked to assign a number to their feelings or attitudes on complex issues. If only they were allowed to do so, they would prefer to express themselves in terms of a range of acceptable positions rather than a particular rating point (cf. Sherif & Hovland, 1961). Part of this indecision may stem from the subject's awareness that his feelings about persons or his attitudes toward issues tend to vacillate over time and to vary somewhat with mood, circumstances, and so on. When asked to estimate a telltale meter, therefore, the subject can in effect tell the truth in many ways. He can tell a truth involving relatively more negative or more positive affect. All that is needed, then, is to make the one further assumption that the person would rather err on the side of being too harsh than too lenient. To take the stereotype study as an example, it may be better to admit being a bigot and have it shown that you are really fairly liberal than to claim tolerance while being revealed as a bigot. The subject therefore hedges his bets, making it less likely for the latter, less preferred, error to occur. On the whole, this is an interesting alternative that is consistent with most of the data thus far collected. However, it does not account for the fact that subjects were more favorable in describing Americans when the meter was employed—unless one assumes that a favorable rating of Americans is an indication of negative attitudes toward others.

A third alternative, not incompatible with those already mentioned, is that the bogus physiological apparatus makes the *affective component of an attitude or evaluation salient over the more cognitive components*. If you ask a person, "What do you think?" he may respond differently than when you ask him, "How do you feel?" Similarly, to ask the subject to predict what his autonomic nervous system is telling the machine is to focus him on the more primitive, affective components of the impression or attitude structure. The resultant response may therefore differ from the verbal rating that is equally honest, but that reflects different dimensions of the attitude being measured. If this were true, it would cast the stereotype findings in quite a different light than if one were to accept the lie detection emphasis of the first alternative. It is easy to jump to the conclusion that the white subjects in the stereotype experiment are really less tolerant and more "racist" than they profess to be. Although this is certainly conceivable, it is equally conceivable that the subjects are trying to cope with two different tasks and responding honestly to both. Regardless of our own particular racial beliefs, we have all heard much about prejudice and probably have well-developed stereotypes about the actuarial likelihood that people have deep-seated negative feelings about other racial groups of which they themselves are unaware. When asked to estimate the affective component of their racial attitudes, therefore, it is not altogether surprising that they would accede to some extent to the implications of this stereotype concerning the nature of prejudice. A similar argument could be developed concerning the actuarial likelihood of hidden negative affect in cases where attraction is being measured. People may have a kind of id–ego conception of their attitudes. They may assume that their ugly, negative feelings are buried in the autonomic nervous system and are expressed only in offguard moments; that portion of the attitude that controls behavior, on the other hand, is tied closely to conscious, rational processes in the intellect.

While it is impossible to choose among these alternatives with our present available knowledge, it is important not to embrace too hastily the proposition that the bogus pipeline is necessarily a better or truer measure of attitude than the standard devices and especially that physiological estimates are a better predictor of overt behavior than questionnaire measures of attitude and attraction. To return to the stereotype study, the results do not necessarily imply that someone who estimates a negative reading to the stimulus Negro will therefore be likely to discriminate against Negroes or to react with hostility to individual blacks. Undoubtedly, different attitudes, and attitudes held by different persons, are differently structured and forecast overt behavior with differential precision. It is quite possible that the more cognitive questionnaire response to an item about race reflects the way a person wants to see himself and the way in which he tries to behave. He may have some reservations about how his autonomic nervous system feels about Negroes, but nevertheless feel that he is in all the more important respects an unprejudiced man. It remains to be seen, certainly, whether the bogus pipeline is any more behavioroid than a questionnaire response, though in many contexts it should be freer of misleading extraneous determinants.

The different alternatives discussed above suggest different research programs. The first alternative, that the bogus pipeline provides a truer measure than other methods of how a person honestly feels, leads to experiments on the extraction of feelings in highly sensitive areas or after interpersonal experiences that arouse unpleasant or embarrassing emotions. The final alternative considers the bogus pipeline as a method tapping a different aspect of attitude. To accept this seriously could lead in the direction of research on the structure of different kinds of attitudes and to an exploration of meaningful discrepancies between cognitive and affective dimensions as a function of issue and setting. It is our belief that research along both of these lines would prove interesting and constructive. By either avenue we shall begin to learn more about the multiple processes of self-attribution and the nebulous phenomenology of emotional experience.

REFERENCES

Aronson, E., & Carlsmith, J. M. Experimentation in social psychology. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology.* (2nd ed.) Vol. II. Reading, Mass.: Addison-Wesley, 1969.

Aronson, E., & Linder, D. E. Gain and loss of esteem as determinants of interpersonal attractiveness. *Journal of Experimental Social Psychology,* 1965, **1**, 156–172.

Behar, L. B. The effects of cognitive dissonance on inappropriate emotional reactions. *Journal of Personality,* 1967, **35**, 505–519.

Bramel, D. A. A dissonance theory approach to defensive projection. *Journal of Abnormal and Social Psychology,* 1962, **64**, 121–129.

Bramel, D. Interpersonal attraction, hostility and perception. In J. S. Mills (Ed.), *Experimental social psychology.* New York: Macmillan, 1969.

Brehm, J. W., & Behar, L. B. Sexual arousal, defensiveness, and sex preferences in affiliation. *Journal of Experimental Research in Personality,* 1966, **1**, 195–200.

Campbell, D. T. The indirect assessment of social attitudes. *Psychological Bulletin,* 1950, **47**, 15–38.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, **56**, 81–105.

Cook, S. W., & Selltiz, C. A multiple-indicator approach to attitude measurement. *Psychological Bulletin,* 1964, **62**, 36–55.

Cooper, J. Personal responsibility and dissonance. The role of foreseen consequences. *Journal of Personality and Social Psychology,* 1971, **18**, 354–363.

Cooper, J. B. Emotion in prejudice. *Science,* 1959, **130**, 314–318.

Crowne, D. P., & Marlowe, D. *The approval motive.* New York: Wiley, 1964.

Edwards, A. L. *The social desirability variable in personality assessment and research.* New York: Dryden, 1957.

Freedman, J., & Fraser, S. C. Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology,* 1966, **4**, 195–202.

Gerard, H. B. Physiological measurement in social psychological research. In P. H. Leiderman & D. Shapiro (Eds.), *Psychobiological approaches to social behavior.* Stanford: Stanford University Press, 1964.

Gerard, H. B. Deviation, conformity, and commitment. In I. D. Steiner & M. Fishbein (Eds.), *Current studies in social psychology.* New York: Holt, Rinehart & Winston, 1965.

Gerard, H. B., & Rabbie, J. M. Fear and social comparison. *Journal of Abnormal and Social Psychology,* 1961, **62**, 586–592.

Goffman, E. *Stigma.* Englewood Cliffs, N. J.: Prentice-Hall, 1963.

Hess, E. H. Attitude and pupil size. *Scientific American,* 1965, **212**, 46–54.

Jones, E. E., Bell, L., & Aronson, E. The reciprocation of attraction from similar and dissimilar others. In C. C. McClintock (Ed.), *Experimental social psychology.* New York: Holt, 1971, in press.

Jones, E. E., & deCharms, R. Changes in social perception as a function of the personal relevance of behavior. *Sociometry,* 1957, **20**, 75–85.

Jones, E. E., Stires, L. K., Shaver, K. G., & Harris, V. A. Evaluation of an ingratiator by target persons and bystanders. *Journal of Personality,* 1968, **36**, 349–385.

Karlins, M., Coffman, T. L., & Walters, G. On the fading of social stereotypes: Studies in three generations of college students. *Journal of Personality and Social Psychology,* 1969, **13**, 1–16.

Lee, C. D. *The instrumental detection of deception.* Springfield, Ill.: Charles C Thomas, 1953.

Leiderman, P. H., & Shapiro, D. (Eds.) *Psychobiological approaches to social behavior.* Stanford: Stanford University Press, 1964.

Lewin, K., Lippitt, R., & White, R. K. Patterns of aggressive behavior in experimentally created social climates. *Journal of Social Psychology,* 1939, **10**, 271–279.

Marlowe, D., Frager, R., & Nuttall, R. L. Commitment to action taken as a consequence of cognitive dissonance. *Journal of Personality and Social Psychology,* 1965, **2**, 864–868.

McGuire, W. J. The nature of attitudes and attitude change. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology.* (2nd ed.) Vol. III. Reading, Mass.: Addison-Wesley, 1969.

Moreno, J. L. *Who shall survive?* Washington, D. C.: Nervous and Mental Disease Publishing Co., 1934.

Orne, M. T. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist,* 1962, **17**, 776–783.

Porier, G. W., & Lott, A. J. Galvanic skin response and prejudice. *Journal of Personality and Social Psychology,* 1967, **5**, 253–259.

Rankin, R. E., & Campbell, D. T. Galvanic skin response to Negro and White experimenters. *Journal of Abnormal and Social Psychology,* 1955, **51**, 30–33.

Riecken, H. W. A program for research on experiments in social psychology. In N. F. Washburne (Ed.), *Decisions, values, and groups.* Vol. 2. New York: Pergamon Press, 1962.

Rosenberg, M. J. When dissonance fails: On eliminating evaluation apprehension from attitude measurement. *Journal of Personality and Social Psychology,* 1965, **1**, 28–43.

Schachter, S. Deviation, rejection, and communication. *Journal of Abnormal and Social Psychology,* 1951, **46**, 190–207.

Scott, W. A. Attitude measurement. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology.* (2nd ed.) Vol. II. Reading, Mass.: Addison-Wesley, 1969.

Shapiro, D., & Crider, A. Psychophysiological approaches in social psychology. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology.* (2nd ed.) Vol. III. Reading, Mass.: Addison-Wesley, 1969.

Sherif, M., & Hovland, C. I. *Social judgment: Assimilation and contrast effects in communication and attitude change.* New Haven: Yale University Press, 1961.

Sigall, H., Aronson, E., & Van Hoose, T. The cooperative subject: Myth or reality? *Journal of Experimental Social Psychology,* 1970, **6**, 1–10.

Sigall, H., & Page, R. Current stereotypes: A little fading, a little faking. *Journal of Personality and Social Psychology,* 1971, **18**, 247–255.

Sternbach, R. A., Gustafson, L. A., & Colier, R. L. Don't trust the lie detector. *Harvard Business Review,* 1962, **40** (6), 127–134.

Valins, S. Cognitive effects of false heart-rate feedback. *Journal of Personality and Social Psychology,* 1966, **4**, 400–408.

Westie, F. R., & DeFleur, M. L. Autonomic responses and their relationship to race attitudes. *Journal of Abnormal and Social Psychology,* 1959, **58**, 340–347.

Zimbardo, P. G., Cohen, A. R., Weisenberg, M., Dworkin, L., & Firestone, I. The control of experimental pain. In P. Zimbardo (Ed.), *The cognitive control of motivation.* Glenview, Ill.: Scott, Foresman, 1969.

(Received April 10, 1970)

Psychological Bulletin
1971, Vol. 76, No. 5, 365–377

MEASURES OF RESPONSE AGREEMENT FOR QUALITATIVE DATA:

SOME GENERALIZATIONS AND ALTERNATIVES

RICHARD J. LIGHT

Harvard University

Various procedures are available for measuring agreement among two or more observers who classify responses among nominal categories. But different problem situations naturally require different measures. Using the general model of a contingency table with fixed margins, this study suggests a measure of level of agreement among several observers when they are compared "internally"; a conditional measure of agreement for several observers compared "internally"; a test for the joint agreement of several observers when their categorizations are compared with a "standard"; and a statistic for evaluating the pattern of agreement between two observers. The first two of these measures are extensions of Cohen's κ statistic. Illustrations are presented for each situation, and results of a Monte Carlo study of the behavior of the pattern agreement statistic are discussed.

Often in psychological experiments two or more observers each assign responses to one of a set of nominal categories. Assume these assignments are made independently. A researcher may then wish to develop a measure of agreement for these observers and to test whether this measure differs significantly from "chance" agreement, or any other prespecified level.

A number of recent studies have suggested techniques for studying observer agreement, and from these studies several issues emerge. First, what is the relationship between measuring association and measuring agreement? Second, should measures of agreement be conditioned upon observed marginal probabilities? Third, can the traditional chi-square approach be fruitfully applied to measuring agreement?

In this study we will consider these issues. We will argue that agreement is a special case of association. We will then distinguish several types of agreement and suggest that each type requires a different sort of measurement procedure. Finally, we will present extensions for and alternatives to techniques currently used to measure agreement. Structuring data for the agreement problem in the format of a contingency table enables us to utilize some well-known results to develop these measures.

If each observer assigns each of n objects to one of C nominal categories, then the contingency table of Figure 1 is specified. The categories on the rows must appear in the same order as the categories on the columns, which permits us to define agreeing items as all those items that appear on the main diagonal of the table, since both observers assign such items to the same category. Goodman and Kruskal (1954) described this case as appropriate for measuring reliability. Note that two rows cannot be permuted unless the corresponding columns are similarly permuted. Thus, the ordering of categories on rows and columns must be identical.

While the layout in Figure 1 is most frequently interpreted as two observers' categorizations of n items, it can be viewed alternatively as each of $2n$ observers' categorizations of a single item. In this latter case, n observers of one type (e.g., mothers) would be paired with n of a second type (e.g., fathers), and each of the $2n$ paired observers assigns an item to one of C categories. For this case, rather than measuring reliability, we might be

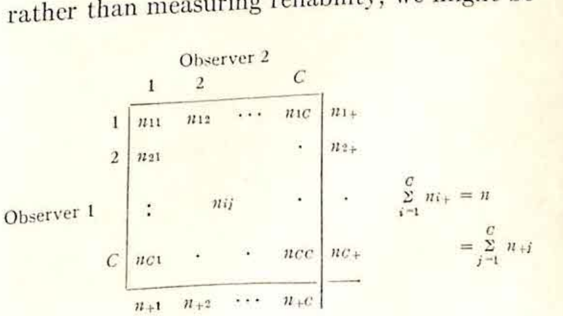


FIG. 1. Notation for two-observer agreement.

---

[1] Requests for reprints should be sent to Richard J. Light, Laboratory of Human Development, Harvard University, Cambridge, Massachusetts 02138.

		Mothers			
		1	2	3	
Fathers	1		50		50
	2			50	50
	3	50			50
		50	50	50	150

FIG. 2. Responses with perfect association, but zero agreement.

interested in estimating an overall level of agreement "within families."

For the special case of $C = 2$ categories, measuring agreement in the resulting 2×2 table is equivalent to measuring association. Thus, chi-square-based measures such as the contingency coefficient and phi, or other measures such as Q, may be used (see Kendall & Stuart, 1967). But when C exceeds 2, the standard measures of association may be inappropriate for measuring agreement.

The distinction between agreement and association for nominal data is that for two or more responses to agree, they must fall into the identical category, while for two responses to be associated requires that the category of one response can be predicted from knowledge of the other's category. Thus, agreement is a special case of association, and a table may exhibit low agreement and high association, as well as high agreement and high association.

To illustrate this distinction, consider Figure 2, a presentation of paired responses of 150 mother-father pairs in answer to a single multiple choice question with nominal alternatives 1, 2, and 3. As only responses on the main diagonal indicate agreement, Figure 2 indicates no agreement; yet, there is perfect association. Given the father's responses, the mother's responses are perfectly predictable, and vice versa, although within-family agreement is zero.

COUNTRY 1

		Mothers		
		1	2	
Fathers	1	60	20	80
	2	20	0	20
		80	20	100

COUNTRY 2

		Mothers		
		1	2	
Fathers	1	20	60	80
	2	0	20	20
		20	80	100

FIG. 3. Example illustrating difference between absolute versus relative agreement.

Note finally that the model underlying Figures 1 and 2 has, in the terminology of Roy and Mitra (1956), both margins variable. Each observer can assign as many of the n items to any category as he wishes. Thus the two observers may have, because of different assignment procedures, substantially asymmetric margins. This possible asymmetry can create difficulties in interpreting agreement. Consider the contrived example in Figure 3.

A single question with two alternatives is presented to 100 mother-father pairs in two countries. Which country has greater within-family agreement? In Country 1, 60% of the pairs agree, while in Country 2 only 40% agree. Yet, given the constraints of the marginals, Country 1's pairs display the least possible agreement, while Country 2's pairs display maximum possible agreement. This problem of disentangling absolute versus relative agreement, noted, for example, by Goodman and Kruskal (1954), suggests the need for a measure conditional on the margins.

Measuring Level of Agreement between Two Observers

Historically much work has been done in developing measures of agreement, or reliability, for continuous data. Ebel (1951), Robinson (1957), Haggard (1958), and Burdock, Fleiss, and Hardesty (1963) have all considered this case. For ranked data, Kendall's W (1962) is well known and frequently used. For qualitative data, however, only in the last few years have some suggestions emerged. Because of the problem (illustrated in Figure 3) of using the raw proportion of agreement $\frac{1}{n}\sum_{i=1}^{C} n_{ii}$, these measures are all conditional in some way on the observed margins.

Cartwright (1956) proposed a measure called alpha, based upon the number of sets of agreeing versus disagreeing pairs of responses, but the required assumption of "equiprobable categories" restricts its use. Another measure of agreement, sensitive to the marginals and which would "correct for chance agreement," was developed by Scott (1955), who proposed:

$$\pi = \frac{p_o - p_e}{1 - p_e}, \qquad [1]$$

where p_o is the observed proportion of agreeing pairs, and p_e is $\sum_{i=1}^{C} p_{ii}$, where p_{ii} is the probability of an entry in cell i, i on the main diagonal of a $C \times C$ table. The p_{ii} were computed, however, on the basis of known population marginal probabilities, and the assumption that both observers had the identical marginal distributions of proportions, equal to the known population values. Cohen (1960) agreed that π was conceptually useful, but noted that actual data often did not meet the assumption of symmetric marginals. He therefore suggested:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \qquad [2]$$

where $p_o = \frac{1}{n} \sum_{i=1}^{C} n_{ii}$ and $p_e = \frac{1}{n^2} \sum_{i=1}^{C} n_{i+}n_{+i}$. κ thus has p_e based upon the observed margins, and requires no assumption of identical marginals.

Note that the κ statistic in Equation 2 takes the general form $1 - \frac{d_o}{d_e}$, where d_o represents observed proportion disagreement, and d_e expected proportion disagreement, under the hypothesis of random agreement. Using this formulation, κ can be expressed as:

$$\kappa = 1 - \left(1 - \frac{\sum_{i=1}^{C} n_{ii}}{n}\right) \Bigg/ \left(1 - \frac{\sum_{i=1}^{C} n_{i+}n_{+i}}{n^2}\right). \qquad [3]$$

Thus κ becomes a ratio of measures of distance, or disagreements, between the two observers, where distances are measured by counting up series of ones and zeros. These distance measures are simply a function of the numbers of agreeing versus disagreeing pairs in the $2n$ total responses.

In a series of further studies, Cohen (1968), Everitt (1968), and Fleiss, Cohen, and Everitt (1969) presented a weighted version of κ, which permits differential weighting of different types of disagreements between observers, as well as large sample standard errors for κ and weighted κ. At present, the simple statistic $\frac{1}{n} \sum_{i=1}^{C} n_{ii}$, which ignores the marginals entirely, κ, and a variation of Cochran's Q statistic proposed by Fleiss (1965) designed for dichotomous responses are the most frequently used agreement statistics. Because of its attractiveness as a simple distance measure of agreement, the κ statistic is extended in the next three sections to measures of conditional agreement and agreement among more than two observers.

A Conditional Measure of Agreement Level—Two Observers

Just as when measuring association in contingency tables it is often useful to measure conditional association, similarly, when measuring agreement, being able to partition an overall κ into a set of partial κ's, κ_{p_i}, $i = 1, \ldots, C$ can contribute additional insight to a data analysis. If the data consist of two observers categorizing n items among C categories, then we might ask about the agreement between the two observers *for only those items which one observer placed into the* i*th specific category*. Or, if the data were drawn from n pairs of observers, such as n mother-father pairs where each of the $2n$ persons assigns a single item to one of C categories, we might ask about the agreement in only those families where the mother selected the ith category.

In some unpublished work, Coleman[2] has used the idea of summing agreeing and disagreeing pairs to develop a conditional version of κ. Taking a $C \times C$ table, a measure of agreement between the two observers for those items which the first observer (appearing on the rows of the table) assigned to the ith specific category is:

$$\kappa_{p_i} = 1 - \left(1 - \frac{n_{ii}}{n_{i+}}\right) \Big/ \left(1 - \frac{n_{+i}}{n}\right). \qquad [4]$$

A similar conditional measure for the second observer (on the columns) could be obtained by merely transposing the columns and rows of a table, and applying Equation 4. Alternatively, n_{i+} and n_{+i} in Equation 4 could be interchanged.

As an illustration of κ_{p_i}, consider a study of 915 father-son pairs. Each of the fathers and each of the sons was asked whether the son or the father was more likely to initiate conversa-

---

[2] Coleman, J. S. Measures of concordance or consensus between members of social groups. Unpublished manuscript, Johns Hopkins University, 1966.

		FATHER'S REPORT		
		Father more	Son more	
SON'S REPORT	Father more	160	65	225
	Son more	348	342	690
		508	407	915

FIG. 4. Report of who initiates conversation.

tions at the dinner table. The results are given in Figure 4. An overall measure of father-son agreement given by κ is:

$$\kappa = 1 - \frac{\left(1 - \frac{502}{915}\right)}{\left(1 - \frac{(225)(508) + (407)(690)}{(915)^2}\right)} = .17.$$

Now we focus on specific families in which the son gave certain responses. For families in which the son reported that the father initiates more, the agreement is:

$$\kappa_{p_1} = 1 - \frac{\left(1 - \frac{160}{225}\right)}{\left(1 - \frac{508}{915}\right)} = .35.$$

The corresponding measure for families in which the son reported that the son initiates more is:

$$\kappa_{p_2} = 1 - \frac{\left(1 - \frac{342}{690}\right)}{\left(1 - \frac{407}{915}\right)} = -.14.$$

Finding a positive κ_{p_1} and a negative κ_{p_2} describes the data very well. Overall, sons tend to report that sons initiate, while fathers predominantly report that fathers initiate. We would expect, therefore, that in families where the son reports the father to be the prime initiator, there would tend to be son-father agreement. In families where the son reported himself to be prime initiator, fathers would be expected to disagree. This is exactly what the κ_{p_i} measures indicated.

To evaluate the statistical significance of κ_{p_i}, as it differs from zero, $\mathrm{Var}_0(\kappa_{p_i})$ must be computed. Extending the results of Fleiss et al. (1969), for the estimated large sample variance of κ, a computationally simple large sample approximation would be:

$$\mathrm{Var}(\kappa_{p_i}) = \frac{p_o(1 - p_o)}{n(1 - p_e)^2}, \qquad [5]$$

where

$$p_o = \frac{n_{ii}}{n_{i+}}, \quad \text{and} \quad p_e = \frac{n_{+i}}{n},$$

which becomes $\mathrm{Var}_0(\kappa_{p_i}) = \dfrac{p_e}{n(1 - p_e)}$ under the null hypothesis of random agreement. For the data in Figure 4, $\mathrm{Var}_0(\kappa_{p_1}) \cong .002$, while $\mathrm{Var}_0(\kappa_{p_2}) \cong .001$. We thus conclude, using the large sample normal approximation, that both κ_{p_1} and κ_{p_2} are different from zero, well beyond the .01 level of significance.

Measuring Level of Agreement Among More Than Two Observers

The κ approach to measuring agreement can be extended further to the situation where more than two observers (say, m observers) each assign n items among C categories. The extension follows directly from the idea of taking pairs of observers' assignment of an item, and seeing whether both observers assign the item to the same category or to different categories. The general form of the multiple agreement statistic is identical to κ's form of $1 - \dfrac{d_o}{d_e}$, where d_o is the observed proportion of disagreements and d_e the expected proportion of disagreements given the m observed marginals. Consider for example the case where $m = 3$. Then the multiple observer agreement statistic κ_m can be written as:

$$\kappa_m = 1 - \left(\frac{A}{n^5}\right) \Big/ \left(\frac{B}{n^6}\right) = 1 - \frac{nA}{B}, \qquad [6]$$

where

$$\begin{aligned} A = {} & \sum_{i \neq j} n_{ij+} \sum_{i \neq l} n_{i++} n_{++l} \sum_{j \neq l} n_{+j+} n_{++l} \\ & + \sum_{i \neq l} n_{i+l} \sum_{i \neq j} n_{i++} n_{+j+} \sum_{j \neq l} n_{+j+} n_{++l} \\ & + \sum_{j \neq l} n_{+jl} \sum_{i \neq j} n_{i++} n_{+j+} \sum_{i \neq l} n_{i++} n_{++l}; \end{aligned}$$

and

$$B = 3 \sum_{i \neq j} n_{i++} n_{+j+} \sum_{i \neq l} n_{i++} n_{++l} \sum_{j \neq l} n_{+j+} n_{++l},$$

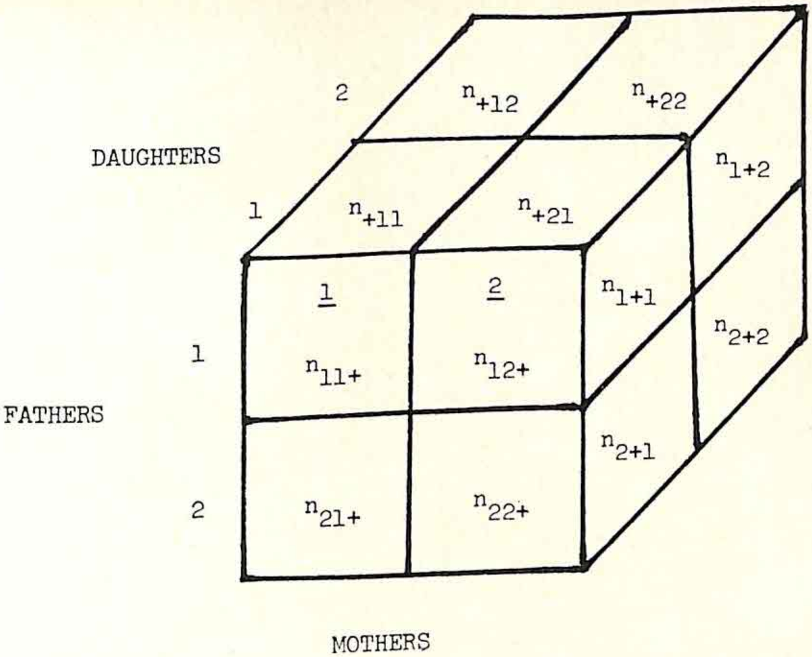


FIG. 5. Notation for measuring agreement among three observers.

where n_{ijl} represents the number of items which the first observer assigns to the ith category, the second observer assigns to the jth category, and the third observer assigns to the lth category.

κ_m can estimate either the joint agreement of m observers categorizing n items, or the agreement among n m-tuplets of observer types (e.g., n father, mother, daughter triplets) categorizing a single item.

An approximate large sample estimate of the variance of κ_m can be found from:

$$\mathrm{Var}(\kappa_m) \cong \frac{p_o(1-p_o)}{n(1-p_e)^2}, \qquad [7]$$

where $p_o = 1 - \dfrac{A}{n^5}$ and $p_e = 1 - \dfrac{B}{n^6}$ with A and B defined in Equation 6. For testing the null hypothesis that $\kappa_m = 0$, of no joint agreement, putting $p_o = p_e$ yields:

$$\mathrm{Var}_0(\kappa_m) \cong \frac{p_e}{n(1-p_e)}, \qquad [8]$$

where p_e is the same as in Equation 7.

To illustrate the computation of κ_m, consider n families, each consisting of a mother, father, and daughter, with each member selecting one of two categories, 1 or 2. The notation can be viewed in Figure 5. The 12 responses in the table are marginal totals.

Figure 6 presents hypothetical data for this example. Note that the within-cell data that result in the marginals of Figure 6 consist of the following eight responses:

$n_{111} = 5$ $\quad n_{211} = 0$
$n_{112} = 5$ $\quad n_{212} = 0$
$n_{121} = 40$ $\quad n_{221} = 5$
$n_{122} = 40$ $\quad n_{222} = 5$

Computing first κ_m, from Equation 6:

$$\begin{aligned} A = {} & (80)(90\cdot 50 + 10\cdot 50)(10\cdot 50 + 90\cdot 50) \\ & + (50)(90\cdot 90 + 10\cdot 10)(10\cdot 50 + 90\cdot 50) \\ & + (50)(90\cdot 90 + 10\cdot 10)(90\cdot 50 + 10\cdot 50) \\ & = 6.1 \times 10^9 \end{aligned}$$

$$B = 3(8200)(5000)(5000) = 6.15 \times 10^{11}.$$

So,

$$\kappa_m = 1 - \frac{nA}{B} = 1 - \frac{100A}{B} \cong .01.$$

To find $\mathrm{Var}_0(\kappa_m)$ from Equation 8:

$$(1 - p_e) = \frac{B}{n^6} = \frac{6.15 \times 10^{11}}{10.0 \times 10^{11}} = .615.$$

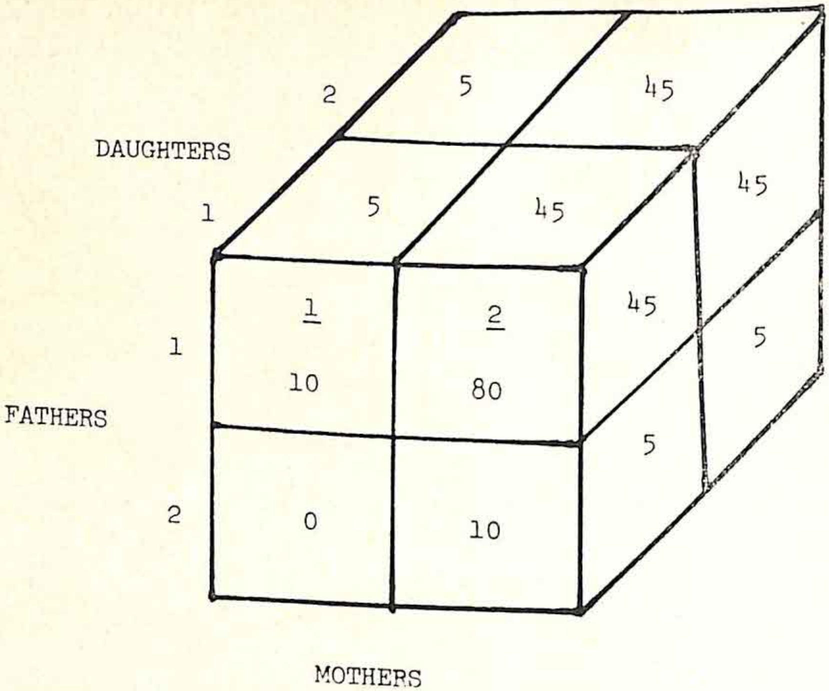


FIG. 6. Set of marginal totals.

Hence,

$$\mathrm{Var}_0(\kappa_m) \cong \frac{.385}{100(.615)} = .0063,$$

which yields an estimated standard error of

$$S_0(\kappa_m) \cong .08.$$

From the large sample normal approximation, $\kappa_m/S_0(\kappa_m) \cong .12$, which leads to the conclusion that $\kappa_m \cong .01$ is not different from zero at any reasonable level of significance.

Conditional Agreement—More Than Two Observers

Just as the conditional measure κ_{p_i} could be defined for two observers, the concept can be extended for $m > 2$. Once again this extension proceeds by summing pairs of agreements and disagreements.

For $m = 3$, a measure of within-family agreement for those families whose "first observer" assigned an item to the ith category (any observer type, e.g., father, can be designated as the "first observer") is:

$$\kappa_{mp_i} = 1 - \frac{n(\sum_j n_{ij+} \sum_l n_{++l} + \sum_j n_{+j+} \sum_l n_{i+l})}{2n_{i++} \sum_j n_{+j+} \sum_l n_{++l}} \quad \begin{matrix} i \neq j \\ i \neq l. \end{matrix} \qquad [9]$$

The large sample approximate variance of κ_{mp_i} can be estimated once again from:

$$\mathrm{Var}(\kappa_{mp_i}) \cong \frac{p_o(1 - p_o)}{n(1 - p_e)^2}, \qquad [10]$$

where now,

$$(1 - p_o) = (\sum_j n_{ij+} \sum_l n_{++l} + \sum_j n_{+j+} \sum_l n_{i+l}) / (n \cdot n_{i++}) \quad i \neq j,\ i \neq l,$$

and

$$(1 - p_e) = \frac{2}{n^2} (\sum_j n_{+j+} \sum_l n_{++l}).$$

To test the hypothesis that $\kappa_{mp_i} = 0$, Equation 8 with p_e from Equation 10 provides a large sample estimated variance.

To illustrate the computation of κ_{mp_i}, we return to the data of Figure 6. The conditional agreement for those families in which the fathers selected the first response is found from Equation 9 to be:

$$\kappa_{mp_1} = 1 - \frac{100[(80)(50) + (90)(45)]}{2(90)(90)(50)}$$

$$= 1 - \frac{8.05 \times 10^5}{8.10 \times 10^5} \cong .01.$$

		Observer 1			
		1	2	3	
Correct assignment	1	$n_{11}^{(1)}$	$n_{12}^{(1)}$	$n_{13}^{(1)}$	n_{1+}
	2	$n_{21}^{(1)}$	$n_{22}^{(1)}$	$n_{23}^{(1)}$	n_{2+}
	3	$n_{31}^{(1)}$	$n_{32}^{(1)}$	$n_{33}^{(1)}$	n_{3+}
		$n_{+1}^{(1)}$	$n_{+2}^{(1)}$	$n_{+3}^{(1)}$	n

		Observer 2			
		1	2	3	
Correct assignment	1	$n_{11}^{(2)}$	$n_{12}^{(2)}$	$n_{13}^{(2)}$	n_{1+}
	2	$n_{21}^{(2)}$	$n_{22}^{(2)}$	$n_{23}^{(2)}$	n_{2+}
	3	$n_{31}^{(2)}$	$n_{32}^{(2)}$	$n_{33}^{(2)}$	n_{3+}
		$n_{+1}^{(2)}$	$n_{+2}^{(2)}$	$n_{+3}^{(2)}$	n

FIG. 7. Data format for comparing joint agreement of several observers with a standard.

Similar computations for the five other κ_{mp_i} yield the following set of results from the data of Figure 6; there are six partial measures, as each of the three family member types chose from two categories.

Partial Agreements

Fathers	Mothers	Daughters
$\kappa_{mp_1} = .01$	$\kappa_{mp_1} = .50$	$\kappa_{mp_1} = -.30$
$\kappa_{mp_2} = .50$	$\kappa_{mp_2} = .01$	$\kappa_{mp_2} = -.30$

We thus see that while for daughters $\kappa_{mp_1} = \kappa_{mp_2}$, for both mothers and fathers $\kappa_{mp_1} \neq \kappa_{mp_2}$. For computation, since Equation 9 gives κ_{mp_i} in terms of a fixed *row* selection (e.g., fathers who choose row one), to find κ_{mp_i} for mothers (and then daughters) it is probably simplest to interchange subscripts of the n_{ijl} in order to view mothers (and then daughters) as being indexed on the rows with the subscript i.

Comparing the Joint Agreement of Several Observers with a Standard

We now focus on a slightly different problem. All of the κ-type measures of agreement have two common properties. First, they assume nothing about a "true, or correct" assignment of items to categories. Responses of two or more observers are not measured relative to any "standard," but rather only with respect to their internal consistency. Second, in the case of more than two observers, κ_m treats all the observers with equal weight. Thus, κ_m might be interpreted as an "overall group agreement" measure.

In some situations, however, neither of these two properties will be appropriate. For some problems, it may be reasonable to assume a "correct set" of responses, to which we wish to compare the responses of one or more observers. For example, when a medical doctor or psychologist believes that he has correctly classified n patients into C symptom categories, he may wish to compare the categorizations of his students, taken as a class, with his "correct" assignments. In other situations, even if we are not willing to assume a set of "correct" assignments, we may be interested in isolating one member of a group as being of special interest, and comparing the responses of the "rest of the group" to those of the isolated member. Thus, even if we do not assume the isolated member's categorizations to be "correct" in an absolute sense, we still use this member's responses as a standard against which to compare the rest of the group. For these situations, κ_m would be inappropriate, as it does not focus on any single group member.

Guetzkow (1950) has suggested several measures to evaluate agreement among observers categorizing items, when each item has a specified "correct" category. He found that agreement among observers when more than four or five categories are involved is primarily determined by the probability of all observers selecting the "correct" category. Tukey (1950) has pointed out, however, that these procedures assume a "response is either correctly classified with one hundred per cent certainty, or it is classified at random [p. 51]," which is an uncommon situation. We therefore consider an alternative procedure.

Assume that a "correctly classified" set of n items, in C categories, is independently categorized by each of m observers. When $m = 1$, then κ serves as a measure of agreement. But when $m > 1$, then κ_m measures the "joint agreement" between the $(m+1)$ sets of categorizations, which treats the correct allocation as "just another observer's alloca-

		Colleague 1			
		Normal	Neurotic	Psychotic	
	Normal	12	6	2	20
Psychologist	Neurotic	4	10	16	30
	Psychotic	24	4	22	50
		40	20	40	100

		Colleague 2			
		Normal	Neurotic	Psychotic	
	Normal	10	3	7	20
Psychologist	Neurotic	14	15	1	30
	Psychotic	6	22	22	50
		30	40	30	100

FIG. 8. Psychologist's categorizations versus each of two junior colleagues' categorizations.

tion." We thus need a test for the joint agreement of the m observers with the "correct" allocation.

Let $n_{ij}{}^{(p)}$ represent the number of responses in the i,j cell of the pth table; $i,j = 1, \ldots, C$; $p = 1, \ldots, m$. For example, suppose $m = 2$ and $C = 3$; then the layout in Figure 7 represents the data. Note that for these two tables in Figure 7, and in general for all m tables, the row margins n_{i+} must be identical for all tables, but the column margins n_{+j} will often differ.

To test the hypothesis of "random agreement" of the group of m observers relative to the "correct" categorization, the statistic G can be used:

$$G = \frac{t_m - E(t_m)}{(\xi t_m)^{1/2}} \sim N(0,1) \qquad [11]$$

where

$$t_m = \sum_{p=1}^{m} \sum_{i=j}^{C} n_{ij}{}^{(p)} \qquad [11a]$$

$$E(t_m) = \frac{1}{n} \sum_{i=j}^{C} [n_{i+}(\sum_{p=1}^{m} n_{+j}{}^{(p)})], \qquad [11b]$$

and

$$\begin{aligned} \xi t_m = {} & \frac{1}{n-1} \sum_{i=j}^{C} [n_{i+}(\sum_{p=1}^{m} n_{+j}{}^{(p)})] \\ & + \frac{1}{n^2(n-1)} Z \\ & - \frac{1}{n(n-1)} \sum_{i=j}^{C} [n_{i+}{}^2(\sum_{p=1}^{m} n_{+j}{}^{(p)})] \\ & - \frac{1}{n(n-1)} \sum_{i=j}^{C} n_{i+}(\sum_{p=1}^{m} (n_{+j}{}^{(p)})^2) \qquad [11c] \end{aligned}$$

		a. Observer 1			
		1	2	3	
	1	4	2	4	10
Observer 2	2	8	4	8	20
	3	8	4	8	20
		20	10	20	50

$\kappa = 0$

		b. Observer 1			
		1	2	3	
	1	2	2	6	10
Observer 2	2	8	6	6	20
	3	10	2	8	20
		20	10	20	50

$\kappa = 0$

FIG. 9. Responses illustrating how patterns of agreement can differ while levels of agreement are identical.

where in Equation 11c above:

$$\begin{aligned} Z = n_{1+}[&n_{1+}\sum_p (n_{+1}{}^{(p)})(n_{+1}{}^{(p)}) \\ &+ n_{2+}\sum_p (n_{+1}{}^{(p)})(n_{+2}{}^{(p)}) + \cdots \\ &+ n_{C+}\sum_p (n_{+1}{}^{(p)})(n_{+C}{}^{(p)})] \\ &+ n_{2+}[n_{1+}\sum_p (n_{+2}{}^{(p)})(n_{+1}{}^{(p)}) \\ &+ n_{2+}\sum_p (n_{+2}{}^{(p)})(n_{+2}{}^{(p)}) + \cdots \\ &+ n_{C+}\sum_p (n_{+2}{}^{(p)})(n_{+C}{}^{(p)})] \\ &\vdots \\ &+ n_{C+}[n_{1+}\sum_p (n_{+C}{}^{(p)})(n_{+1}{}^{(p)}) \\ &+ n_{2+}\sum_p (n_{+C}{}^{(p)})(n_{+2}{}^{(p)}) + \cdots \\ &+ n_{C+}\sum_p (n_{+C}{}^{(p)})(n_{+C}{}^{(p)})]. \end{aligned}$$

G compares the observed agreements of each of the m respondents with the "correct" assignment, while recomputing the expected number of agreements for each respondent based upon his margin assignments $n_{+1}{}^{(p)}, \ldots, n_{+C}{}^{(p)}$, $p = 1, \ldots, m$. G is developed from the well-known results that for any contingency table, given fixed marginals:

$$E(n_{ij}) = \frac{n_{i+}n_{+j}}{n}$$

and

$$\text{Var}(n_{ij}) = \frac{n_{i+}(n_{i+} - 1)n_{+j}(n_{+j} - 1)}{n(n - 1)}$$

and for any two cells not in the same row or same column:

$$\text{Cov}(n_{ij}, n_{gh}) = \frac{n_{i+}n_{g+}n_{+j}n_{+h}}{n^2(n - 1)} \quad i \neq g,\ j \neq h.$$

The variance of the sum of the entries along the agreement diagonal, for a single observer, is:

$$\sum_{i=j}^{C} \text{Var}(n_{ij}) + 2 \sum_{i \neq g}^{C} \text{Cov}(n_{ii}, n_{gg}).$$

Summing these single observer variances over the m observers, together with the pairwise covariances over the m tables, yields the total variance ξ_{t_m}. For testing, G will be approximately normally distributed around $E(t_m)$ under the null hypothesis of "random agreement."

As an example of the use of G, suppose a psychologist classifies each of 100 subjects into one of three classes; normal, neurotic, or psychotic. Suppose further that each of two junior colleagues independently categorizes the same subjects, and we wish to see whether the joint assignments made by the two colleagues "agree beyond chance" with the ratings of the psychologist. The data appear in Figure 8. Computing t_m, $E(t_m)$, and ξ_{t_m} from the formulas in Equation 11: From Equation 11a, $t_m = 91$; from Equation 11b, $E(t_m) = 67$; from Equation 11c,

$$\xi_{t_m} = \frac{1}{99}(6{,}700) + \frac{1}{990{,}000}(22{,}450{,}000) - \frac{1}{9{,}900}(257{,}000) - \frac{1}{9{,}900}(235{,}000) = 40.7.$$

A test of the hypothesis of random agreement is given by:

$$G = \frac{91 - 67}{(40.7)^{1/2}} = \frac{24.0}{6.4} = 3.75.$$

Whether to use a one-sided or two-sided test in this problem is up to a particular researcher, but in either case, the value $G = 3.75$ is significant beyond the .01 level, and we may conclude that the joint categorizations of the two junior colleagues agreed with the psychologist's categorization significantly more than we would have expected due to chance.

Distinguishing Pattern from Level of Agreement between Two Observers

We return finally to the case of $m = 2$ observers. The statistic κ gives a measure of how the observed *overall level* of agreement compares with the expected overall level. It is completely insensitive to differences between the observed and expected *patterns* of agreement. For example, both of the 3×3 tables in Figure 9 have $\kappa = 0$, and identical margins. However, while the observed *levels* of agreement displayed by both Figure 9a and Figure 9b are identical to those expected by chance, the *patterns* of agreement, which contribute to the overall level, are different for the two tables. Figure 9a has the agreements distributed exactly as ex-

a. Observer 1

		1	2	3	
	1	4	2	4	10
Observer 2	2	8	4	8	20
	3	8	4	8	20
		20	10	20	50

$\chi^2 = 0$

b. Observer 1

		1	2	3	
	1	4	5	1	10
Observer 2	2	5	4	11	20
	3	11	1	8	20
		20	10	20	50

$\chi^2 = 12.38$

FIG. 10. Data illustrating why standard chi-square statistic is inappropriate for measuring agreement.

pected due to chance, on the agreement diagonal. Figure 9b has the agreements distributed *differently* from expectation. Yet κ for both figures is zero.

An agreement measure that is sensitive to deviations between observed and expected counts *in particular cells* would thus be useful. Cohen (1968) noted this need when pointing out that the standard chi-square approach to measuring association has sometimes been used to measure agreement, although it is inappropriate because of its sensitivity to discrepancies between observed and expected cell counts in disagreement (off-diagonal) cells of the table. For example, in Figure 10, both Figure 10a and Figure 10b have the identical levels *and* patterns of agreement, yet in Figure 10a chi-square shows exactly "chance" agreement, while in Figure 10b chi-square registers a large discrepancy between expected and observed cell counts. The problem with using the standard chi-square, then, is its sensitivity to discrepancies between observed and expected counts in *every* cell of a table. To examine the pattern of agreement between two observers, we need a statistic sensitive only to discrepancies on the agreement diagonal.

A modified chi-square approach enables us to examine pattern agreement. Lumping together all the off-diagonal cells, so that discrepancies between observed and expected cell counts will not affect our measure, leads to the statistic:

$$A_p = \sum_{i=j}^{C}\sum^{C} \frac{\left(n_{ij} - \dfrac{n_{i+}n_{+j}}{n}\right)^2}{\dfrac{n_{i+}n_{+j}}{n}} + \frac{1}{\sum\limits_{i\neq j}^{C}\sum\limits^{C} \dfrac{n_{i+}n_{+j}}{n}}\left(\sum_{i\neq j}^{C}\sum^{C} n_{ij} - \sum_{i\neq j}^{C}\sum^{C} \frac{n_{i+}n_{+j}}{n}\right)^2. \quad [12]$$

This measure of pattern agreement A_p is merely a variant of chi-square that focuses on the agreement cells individually, and pools all disagreement cells into a set. When used together with κ, which estimates the level of observer agreement, these two measures should offer a more complete picture of the relationship between two observers than either measure could give individually. Under random agreement, A_p is distributed asymptotically as chi-square with C degrees of freedom, for a $C \times C$ table.

To illustrate A_p, return to the data in Figure 9. In Figure 9a, $A_p = 0$, as each cell on the agreement diagonal and the total of all nonagreement cells have identical observed and expected cell counts. For Figure 9b,

$$A_p = \frac{(2-4)^2}{4} + \frac{(6-4)^2}{4} + \frac{(8-8)^2}{8} + \frac{(34-34)^2}{34} = 2.0,$$

and a chi-square value of 2.0 with 3 degrees of freedom does not indicate that the observed agreement pattern differs from the expected pattern at, say, the .05 level. Thus, for Figure 9b, we cannot conclude from $\kappa = 0$ and $A_p = 2.0$ that either the level or the pattern of agreement is different from the expected level and pattern.

Note the possibility that κ can be zero while A_p is very large and highly significant; whereas if A_p is zero, κ is constrained to be zero also.

Report of Some Sampling Experiments on A_p

Cochran (1954) and others, for example, Lewontin and Felsenstein (1965), have pointed out that for the standard chi-square test in contingency table, the presence of cells whos $E(n_{ij})$ is very small, such as 1 or 2, may severel impair the chi-square statistic's behavior as an approximation of its asymptotic reference distribution. A_p, while asymptotically distributed as chi-square with C degrees of freedom, is derived from the general model of the $(C + 1)$-nomial. That is, while A_p is computed from the

TABLE 1

MONTE CARLO STUDIES OF PATTERN AGREEMENT STATISTIC A_p

Second marginal	Mean and variance	First marginal									
A. 3 × 3 Contingency tables											
		.33 .33 .34		.40 .40 .20		.20 .30 .50		.20 .20 .60		.80 .10 .10	
		$n =$ 50	200	$n =$ 50	200	$n =$ 50	200	$n =$ 50	200	$n =$ 50	200
.33 .33 .34	M	2.97	2.97	2.99	3.01	3.02	2.96	3.02	3.01	3.01	3.01
	Var	6.25	5.82	6.47	5.95	6.18	6.11	6.38	6.31	7.46	6.31
.40 .40 .20	M			3.03	2.97	3.06	2.97	2.98	2.95	2.94	3.06
	Var			6.29	6.00	6.73	5.99	6.13	6.10	6.80	6.25
.20 .30 .50	M					3.00	3.04	2.97	3.00	3.01	3.03
	Var					6.37	6.25	6.59	6.09	6.61	6.49
.20 .20 .60	M							3.03	2.99	3.04	3.06
	Var							6.71	6.30	7.23	6.30
.80 .10 .10	M									2.96	2.98
	Var									9.49	7.03
B. 4 × 4 Contingency tables											
		.25 .25 .25 .25		.30 .30 .20 .20		.10 .10 .40 .40		.10 .20 .30 .40		.70 .10 .10 .10	
		$n =$ 50	200	$n =$ 50	200	$n =$ 50	200	$n =$ 50	200	$n =$ 50	200
.25 .25 .25 .25	M	3.91	4.01	4.05	4.01	3.99	4.01	4.00	4.04	3.99	3.99
	Var	8.34	8.21	9.10	8.19	9.09	8.53	8.74	8.64	9.11	8.38
.30 .30 .20 .20	M			3.99	4.01	4.04	4.01	4.01	3.98	4.03	3.97
	Var			8.77	8.01	10.07	8.79	8.86	8.21	9.99	8.18
.10 .10 .40 .40	M					3.97	3.92	3.95	4.00	4.01	3.96
	Var					10.76	8.48	10.05	8.38	10.59	8.41
.10 .20 .30 .40	M							4.01	4.01	4.02	4.08
	Var							9.79	9.02	10.33	8.70
.70 .10 .10 .10	M									4.01	3.97
	Var									13.86	8.57

Note.—The table gives the mean and variance of A_p for various combinations of marginal probability vectors, for $n = 50$ and $n = 200$ under the null hypothesis.

contingency table format, it is simply a function of the individual cell probabilities on the C diagonal cells.

We therefore present the results of some sampling experiments, which enable us to study how well A_p is referenced by chi-square (C) when (*a*) one or both observers assign items to categories with very unevenly distributed margins, such as with probabilities .80, .10, .10 in a three-category assignment; (*b*) the probability structure of the two margins are substantially different, due to differences in overall rating strategies between the two observers; (*c*) because of the marginal probability structures, one or more $E(n_{ij})$ on the agreement diagonal is very small.

Table 1 gives the results of the sampling experiments; Table 1A for 3×3 tables, and Table 1B for 4×4's. In these experiments, the probability structures of both marginals were specified, together with the total number of items n to be categorized by each observer. Five marginal probability structures were examined for each observer, which when combined in pairs and with themselves gave a total of 15 possible contingency tables. Fifteen 3×3 tables and fifteen 4×4 tables were thus examined. For each of these 30 tables, samples were generated with $n = 50$ observations, to represent a small total sample, and $n = 200$ observations, to represent a moderately large sample. Thus, 60 different experiments were run.

For each experiment, 5,000 samples were generated. A_p was computed for each sample, and the mean and variance from each set of 5,000 samples are given in Table 1A and Table 1B. In Table 1A, the expected mean and variance for A_p under the asymptotic reference distribution are 3.00 and 6.00, respectively; in Table 1B, these are 4.00 and 8.00. Examining the results of these experiments helps us to analyze the behavior of A_p under the three conditions posed above.

Results

First, regardless of the probability structure of the table, A_p has its observed variance closer to the expected values when $n = 200$ than when $n = 50$. The mean always fits well. This was true for all 30 table structures. Thus, the larger the total number of items to be categorized, the better A_p will be approximated by chi-square (C).

Second, for 3×3 tables, $n = 50$ results in A_p being quite well approximated by chi-square's first two moments for all probability structures except when one or both observers concentrate most of the items into a single category. When this happens, the actual variance of A_p is substantially underestimated by chi-square (C). For example, when both observers tend to assign 80% of the items to one category, and 10% to each of the other two categories, the variance of A_p for $n = 50$ is found to be approximately 9.49 rather than the expected 6.00.

Third, for 4×4 tables, when $n = 50$, chi-square (4) will generally underestimate the actual variance of A_p. Only when both observers assign items to categories with approximately equal probabilities (a probability of .25 for each category, for both observers), is the experimentally observed variance (8.34) approximately equal to the expected 8.00.

Fourth, for all 30 tables, $n = 200$ yielded a mean and variance for A_p reasonably close to their expected values. Thus, for most marginal probability structures, a sample size of $n = 200$ for 3×3 or 4×4 tables indicates chi-square (C) will be a reasonable reference distribution.

Fifth, for table structures that have cells on the agreement diagonal with low $E(n_{ij})$, A_p has an experimental variance somewhat larger than the expected variance. Thus, for any given n, the larger the probabilities in each of the cells on the agreement diagonal, the better the referencing of A_p by chi-square (C).

Finally, because the actual variance of A_p generally exceeds the expected variance from chi-square, the use of chi-square to examine a table for pattern agreement will result in a *conservative* test. Researchers may wish to use this knowledge when interpreting probability levels.

REFERENCES

Burdock, E. I., Fleiss, J. L., & Hardesty, A. S. A new view of interobserver agreement. *Personnel Psychology*, 1963, **16**, 373–384.

Cartwright, D. S. A rapid non-parametric estimate of multi-judge reliability. *Psychometrika*, 1956, **21**, 17–29.

Cochran, W. G. Some methods for strengthening the common chi-square tests. *Biometrics*, 1954, **10**, 417–451.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, **20**, 37–46.

Cohen, J., Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 1968, **70**, 213–220.

Ebel, R. L. Estimation of the reliability of ratings. *Psychometrika*, 1951, **16**, 407–424.

Everitt, B. S. Moments of the statistics κ and weighted kappa. *British Journal of Mathematical and Statistical Psychology*, 1968, **21**, 97–103.

Fleiss, J. L. Estimating the accuracy of dichotomous judgments. *Psychometrika*, 1965, **30**, 469–479.

Fleiss, J. L., Cohen, J., & Everitt, B. S. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 1969, **72**, 323–327.

Guetzkow, H. Unitizing and categorizing problems in coding qualitative data. *Journal of Clinical Psychology*, 1950, **6**, 47–58.

Goodman, L. A., & Kruskal, W. H. Measures of association for cross classifications. *Journal of the Americal Statistical Association*, 1954, **49**, 732–764.

Haggard, E. A. *Intraclass correlation and the analysis of variance*. New York: Dryden Press, 1958.

Kendall, M. G. *Rank correlation methods*. (3rd ed.) New York: Hafner, 1962.

Kendall, M. G., & Stuart, A. *The advanced theory of statistics*. (2nd ed.) Vol. II. New York: Hafner, 1967.

Lewontin, R. C., & Felsenstein, J. The robustness of homogeneity tests in $2 \times n$ tables. *Biometrics*, 1965, **21**, 19–33.

Roy, S. N., & Mitra, S. K. An introduction to some non-parametric generalizations of analysis of variance and multivariate analysis. *Biometrika*, 1956, **43**, 361–376.

Robinson, W. S. The statistical measurement of agreement. *American Sociological Review*, 1957, **22**, 17–27.

Scott, W. A. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 1955, **19**, 321–325.

Tukey, J. W., Discussion on Symposium. *Journal of Clinical Psychology*, 1950, **6**, 61–74.

(Received April 7, 1970)

Manuscripts Accepted for Publication in the

Psychological Bulletin

Comment on "Equivalence of Information in Concept Identification": Leonard M. Giambra*: Department of Psychology, Miami University, Oxford, Ohio 45056.

Nonparametric Indexes of Sensitivity and Response Bias: John T. E. Richardson*: 84 Penlands Road, Haywards Heath, Sussex, England.

Relation between Signal Detectability Theory and the Traditional Procedures for Measuring Thresholds: An Addendum: Michael Treisman*: Department of Psychology, University of Reading, Building 3, Earley Gate, White Knights, Reading, England.

---

* Asterisk indicates author for whom the address is supplied.

Psychological Bulletin
1971, Vol. 76, No. 5, 378–382

MEASURING NOMINAL SCALE AGREEMENT AMONG MANY RATERS[1]

JOSEPH L. FLEISS[2]

Biometrics Research, New York State Department of Mental Hygiene

The statistic kappa was introduced to measure nominal scale agreement between a fixed pair of raters. In this paper kappa is generalized to the case where each of a sample of subjects is rated on a nominal scale by the same number of raters, but where the raters rating one subject are not necessarily the same as those rating another. Large sample standard errors are derived, and a numerical example is given.

Cohen introduced the statistics kappa (Cohen, 1960) and weighted kappa (Cohen, 1968) to measure the degree of agreement between two raters who rate each of a sample of subjects on a nominal scale. Both kappa and weighted kappa incorporate a correction for the extent of agreement expected by chance alone. Kappa is useful when all disagreements may be considered equally serious, and weighted kappa is useful when the relative seriousness of the different kinds of disagreement can be specified. Properties of these two statistics have been studied by Everitt (1968) and by Fleiss, Cohen, and Everitt (1969).

The use of kappa and weighted kappa is restricted to the case both where the number of raters is two and where the same two raters rate each subject. Generalizations are therefore necessary for the case of more than two raters and for the case where the raters judging one subject are not necessarily the same as those judging another.

In this paper we consider only the generalization of unweighted kappa to the measurement of agreement among any constant number of raters where there is no connection between the raters judging the various subjects. An important special case is when subjects are rated by different *pairs* of raters. We do not consider the case of varying numbers of raters, or the case where each subject is rated by the same group of more than two raters, or the case where weights can be assigned to the various disagreements.

NOTATION

Let N represent the total number of subjects, n the number of ratings per subject, and k the number of categories into which assignments are made. Let the subscript i, where $i = 1, \ldots, N$, represent the subjects, and the subscript j, where $j = 1, \ldots, k$, represent the categories of the scale.

Define n_{ij} to be the number of raters who assigned the ith subject to the jth category, and define

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}. \qquad [1]$$

The quantity p_j is the proportion of all assignments which were to the jth category. Since $\Sigma_j n_{ij} = n$, therefore $\Sigma_j p_j = 1$.

Table 1 presents data on which of $k = 5$ diagnoses were given each of $N = 30$ patients by $n = 6$ psychiatrists. Thus, $n_{14} = 6$ and $n_{1j} = 0$ for $j \neq 4$; $n_{22} = 3$, $n_{25} = 3$, and $n_{2j} = 0$ for $j \neq 2$ and $j \neq 5$; etc. The overall proportions are $p_1 = .144$, $p_2 = .144$, $p_3 = .167$, $p_4 = .306$, and $p_5 = .239$. A total of 43 psychiatrists provided diagnoses. In the actual study (Sandifer, Hordern, Timbury, & Green, 1968), between 6 and 10 psychiatrists from the pool of 43 were unsystematically selected to diagnose a subject. For the illustrative purposes of this paper, randomly selected diagnoses were dropped to bring the number of assignments per patient down to a constant of six.

---

[1] This work was supported in part by Grant MH 08534 and in part by Grant MH 09191, both from the National Institute of Mental Health. Myron G. Sandifer, Jr., Associate Dean for Academic Affairs of the University of Kentucky College of Medicine, kindly provided the data on which Table 1 is based.

[2] Also at Columbia University. Requests for reprints should be sent to Joseph L. Fleiss, Biometrics Research, State of New York Department of Mental Hygiene, 722 West 168 Street, New York, N. Y. 10032.

MEASURING OVERALL AGREEMENT

The extent of agreement among the n raters for the ith subject may be indexed by the proportion of agreeing pairs out of all the $n(n-1)$ possible pairs of assignments. This proportion is, say,

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij}-1)$$

$$= \frac{1}{n(n-1)} \left(\sum_{j=1}^{k} n_{ij}^2 - n\right). \quad [2]$$

Thus, $P_1 = 1$; $P_2 = .40$; etc. The overall extent of agreement may then be measured by the mean of the P_is, say,

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i$$

$$= \frac{1}{Nn(n-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}^2 - Nn\right). \quad [3]$$

For the data of Table 1,

$$\bar{P} = \frac{1}{(30)(6)(5)}(680 - 30 \times 6) = .5556. \quad [4]$$

The value $\bar{P} = .5556$ may be interpreted as follows. Let a subject be selected at random and diagnosed by a randomly selected psychiatrist. If the subject were also diagnosed by a second randomly selected psychiatrist, the second diagnosis would agree with the first over 55% of the time. Measures like $\bar{P}$ have been used to measure diagnostic agreement by Sandifer, Pettus, and Quade (1964), by Sandifer, Hordern, Timbury, and Green (1968) and by Sandifer, Fleiss, and Green (1968).

It is clear, however, that some degree of agreement is to be expected solely on the basis of chance. In fact, if the raters made their assignments purely at random, one would expect the mean proportion of agreement to be

$$\bar{P}_e = \sum_{j=1}^{k} p_j^2. \quad [5]$$

For the data of Table 1,

$$\bar{P}_e = .144^2 + .144^2 + \cdots + .239^2 = .2201. \quad [6]$$

TABLE 1

DIAGNOSES ON 30 SUBJECTS BY SIX RATERS PER SUBJECT

Subject	Category Depression ($j=1$)	Personality disorder ($j=2$)	Schizophrenia ($j=3$)	Neurosis ($j=4$)	Other ($j=5$)
1				6	
2		3			3
3		1	4		1
4					6
5		3		3	
6	2		4		
7			4		2
8	2		3	1	
9	2			4	
10					6
11	1			5	
12	1	1		4	
13		3	3		
14	1			5	
15		2		3	1
16			5		1
17	3			1	2
18	5	1			
19		2		4	
20	1		2		3
21					6
22		1		5	
23		2		1	3
24	2			4	
25	1			4	1
26		5		1	
27	4				2
28		2		4	
29	1		5		
30					6
Total	26	26	30	55	43
p_j	.144	.144	.167	.306	.239

The quantity $1 - \bar{P}_e$ measures the degree of agreement attainable over and above what would be predicted by chance. The degree of agreement actually attained in excess of chance is $\bar{P} - \bar{P}_e$, so that a normalized measure of overall agreement, corrected for the amount expected by chance, is

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}. \quad [7]$$

For the data of Table 1,

$$\kappa = \frac{.5556 - .2201}{1 - .2201} = .430. \quad [8]$$

An equivalent expression for κ is

$$\kappa = \frac{\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}^2 - Nn[1 + (n-1) \sum_{j=1}^{k} p_j^2]}{Nn(n-1)(1 - \sum_{j=1}^{k} p_j^2)}. \quad [9]$$

STANDARD ERROR OF κ

It seems reasonable to interpret the absence of real agreement among the n raters as their inability to distinguish one subject from another. Such an inability implies that the raters apply the overall rates of assignment, $\{p_j\}$, to each and every subject. This in turn implies that under the hypothesis of no agreement, the n assignments on each subject are multinomial variables with probabilities $p_1, \ldots, p_k$.

If N, the number of subjects, is large, we may assume that the proportions $p_1, \ldots, p_k$ given by Equation 1 are constants so that by Equation 9, κ is essentially a function only of the random variable $\Sigma_i \Sigma_j n_{ij}^2$. We assume further that even though $p_1, \ldots, p_k$ are held constant, N is large enough so that the sum of the squared numbers of assignments for the ith subject, $\Sigma_j n_{ij}^2$, is independent of that for the i'th subject, $\Sigma_j n_{i'j}^2$. Thus, if N is large, the variance of κ under the hypothesis of no agreement beyond chance is approximately equal to

$$\text{Var}(\kappa) = \frac{\sum_{i=1}^{N} \text{Var}(\sum_{j=1}^{k} n_{ij}^2)}{N^2 n^2 (n-1)^2 (1 - \sum_{j=1}^{k} p_j^2)^2}. \quad [10]$$

Now,

$$\begin{aligned} \text{Var}(\sum_j n_{ij}^2) &= E(\sum_j n_{ij}^2)^2 - [E(\sum_j n_{ij}^2)]^2 \\ &= E(\sum_j n_{ij}^4) + E(\sum_{j \neq m} \sum n_{ij}^2 n_{im}^2) \\ &\quad - [E(\sum_j n_{ij}^2)]^2. \quad [11] \end{aligned}$$

The required moments are found to be

$$\begin{aligned} E(\sum_j n_{ij}^4) &= n + n(n-1) \\ &\quad \times [7 \sum_j p_j^2 + 6(n-2) \sum_j p_j^3 \\ &\quad + (n-2)(n-3) \sum_j p_j^4], \quad [12] \end{aligned}$$

$$\begin{aligned} E(\sum_{j \neq m} \sum n_{ij}^2 n_{im}^2) &= n(n-1) \\ &\quad + n(n-1)[(2n-5) \sum_j p_j^2 \\ &\quad + (n-2)(n-3)(\sum_j p_j^2)^2 - 2(n-2) \sum_j p_j^3 \\ &\quad - (n-2)(n-3) \sum_j p_j^4], \quad [13] \end{aligned}$$

and

$$\begin{aligned} [E(\sum_j n_{ij})]^2 &= n^2 + n(n-1) \\ &\quad \times [2n \sum_j p_j^2 + n(n-1)(\sum_j p_j^2)^2]. \quad [14] \end{aligned}$$

Thus,

$$\begin{aligned} &\text{Var}(\sum_j n_{ij}^2) \\ &= 2n(n-1)[\sum_j p_j^2 - (2n-3)(\sum_j p_j^2)^2 \\ &\quad + 2(n-2) \sum_j p_j^3], \quad [15] \end{aligned}$$

and, because $\Sigma_i \text{Var}(\Sigma_j n_{ij}^2) = N \text{Var}(\Sigma_j n_{ij}^2)$, therefore

$$\begin{aligned} \text{Var}(\kappa) &= \frac{2}{Nn(n-1)} \\ &\quad \times \frac{\sum_j p_j^2 - (2n-3)(\sum_j p_j^2)^2 + 2(n-2) \sum_j p_j^3}{(1 - \sum_j p_j^2)^2}. \quad [16] \end{aligned}$$

For the data of Table 1, we have already found that $\Sigma_j p_j^2 = .2201$. Further, $(\Sigma_j p_j^2)^2 = .0484$ and $\Sigma_j p_j^3 = .0529$. Thus, the variance of κ is approximately

$$\begin{aligned} \text{Var}(\kappa) &= \frac{2}{(30)(6)(5)} \\ &\quad \times \frac{.2201 - (9)(.0484) + (2)(4)(.0529)}{(1 - .2201)^2} \\ &= .000759, \quad [17] \end{aligned}$$

and the standard error (SE) of κ is approximately

$$\text{SE}(\kappa) = .028. \quad [18]$$

Under the hypothesis of no agreement beyond chance, $\kappa / \text{SE}(\kappa)$ will, by the central limit theorem, be approximately distributed as a standard normal variate. In our case,

$$\frac{\kappa}{\mathrm{SE}(\kappa)} = \frac{.430}{.028} = 15.4, \qquad [19]$$

so we would infer that the overall agreement in assignment to the k categories is significantly greater than chance.

Agreement on a Particular Category

We consider now the extent of agreement in assigning a subject to category j. To motivate the choice of the basic index of agreement, we suppose that each subject is assigned to one of the k categories by a randomly selected rater. Let only those subjects assigned to category j be rated again, now by another randomly selected rater. The index is the conditional probability that the second assignment is to category j, given that the first was to category j. It is calculated as

$$\bar{P}_j = \frac{\sum_{i=1}^{N} n_{ij}(n_{ij}-1)}{\sum_{i=1}^{N} n_{ij}(n-1)} = \frac{\sum_{i=1}^{N} n_{ij}^2 - Nnp_j}{Nn(n-1)p_j}, \qquad [20]$$

where p_j is defined by Equation 1.

Under the hypothesis of no agreement beyond chance, we would expect this probability to equal the unconditional probability of an assignment to category j, namely p_j. A measure of the extent of agreement beyond chance in assignment to category j is then

$$\kappa_j = \frac{\bar{P}_j - p_j}{1-p_j}$$

$$= \frac{\sum_{i=1}^{N} n_{ij}^2 - Nnp_j[1+(n-1)p_j]}{Nn(n-1)p_jq_j}, \qquad [21]$$

where $q_j = 1 - p_j$. Note that κ, the measure of overall agreement (Equation 9), is a weighted average of the κ_js. In fact,

$$\kappa = \Sigma_j p_j q_j \kappa_j / \Sigma_j p_j q_j.$$

If N is large, we can assume that p_j is a constant and, even with this restriction, that n_{ij}^2 is independent of $n_{i'j}^2$ for $i \neq i'$. Under the hypothesis of no more than chance agreement in assignment to category j, the quantity n_{ij} has a binomial distribution with parameters n and p_j.

TABLE 2

Statistics for Measuring Agreement on Each of the Five Categories of Table 1

Category	$\Sigma_i n_{ij}^2$	p_j	$\bar{P}_j$	κ_j	Var (κ_j)	$\kappa_j/\mathrm{SE}(\kappa_j)$
1	72	.144	.356	.248	.0130	2.17
2	72	.144	.356	.248	.0130	2.17
3	120	.167	.598	.517	.0136	4.44
4	229	.306	.632	.470	.0195	3.36
5	187	.239	.669	.565	.0163	4.43

Under this hypothesis,

$$\mathrm{Var}(n_{ij}^2) = np_jq_j \times [(1+2(n-1)p_j)^2 + 2(n-1)p_jq_j]. \qquad [22]$$

Therefore, the approximate variance of κ_j is

$$\mathrm{Var}(\kappa_j) = \frac{[1+2(n-1)p_j]^2 + 2(n-1)p_jq_j}{Nn(n-1)^2p_jq_j}. \qquad [23]$$

The hypothesis may be tested by referring the quantity $\kappa_j/\mathrm{SE}(\kappa_j) = \kappa_j/\sqrt{\mathrm{Var}(\kappa_j)}$ to tables of the standard normal distribution.

Table 2 gives summary statistics for each of the five categories. The calculations for Category 1 (consisting of involutional melancholia, manic-depressive depression, and psychotic depression) are illustrated in detail. Since $\Sigma_i n_{i1}^2 = 72$, and $p_1 = .144$, the basic index is, by Equation 20,

$$\bar{P}_1 = \frac{72 - (30)(6)(.144)}{(30)(6)(5)(.144)} = .356, \qquad [24]$$

indicating that, given that one assignment of a subject is to Category 1, the chances of a second one being to the same category are over one in three.

By Equation 21,

$$\kappa_1 = \frac{.356 - .144}{1 - .144} = .248. \qquad [25]$$

The approximate variance of κ_1 is, by Equation 23,

$$\mathrm{Var}(\kappa_1) = \frac{[1+(2)(5)(.144)]^2 + (2)(5)(.144)(.856)}{(30)(6)(5)^2(.144)(.856)} = .0130, \qquad [26]$$

so that the approximate standard error is .114. Thus,

$$\frac{\kappa_1}{\mathrm{SE}(\kappa_1)} = \frac{.248}{.114} = 2.17, \qquad [27]$$

indicating that the agreement in assignment to Category 1 is significantly better than chance at the .05 level.

REFERENCES

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, **20**, 37–46.

Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 1968, **70**, 213–220.

Everitt, B. S. Moments of the statistics kappa and weighted kappa. *British Journal of Mathematical and Statistical Psychology*, 1968, **21**, 97–103.

Fleiss, J. L., Cohen, J., & Everitt, B. S. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 1969, **72**, 323–327.

Sandifer, M. G., Fleiss, J. L., & Green, L. M. Sample selection by diagnosis in clinical drug evaluations. *Psychopharmacologia*, 1968, **13**, 118–128.

Sandifer, M. G., Hordern, A., Timbury, G. C., & Green, L. M. Psychiatric diagnosis: A comparative study in North Carolina, London and Glasgow. *British Journal of Psychiatry*, 1968, **114**, 1–9.

Sandifer, M. G., Pettus, C., & Quade, D. A study of psychiatric diagnosis. *Journal of Nervous an Mental Diseases*, 1964, **139**, 350–356.

(Received May 12, 1970)

ERRATUM

In the article "Current Research on the Frequency of Dream Recall" by David Cohen in the June 1970 issue, lines 26–30 of the left column of page 437 should read as follows: "Relatively high anxiety was related to significantly lower frequency of dream recall in groups marked by intellectualization and abnegation. . . ."

VOL. 76, NO. 6 DECEMBER 1971

Psychological Bulletin

PROPRIOCEPTION AND THE TIMING OF MOTOR RESPONSES

RICHARD A. SCHMIDT[1]

Department of Physical Education, University of Michigan

Two hypotheses that relate motor response timing to proprioceptive feedback are presented, along with specific predictions. The input hypothesis states that the subject uses the inflow of proprioceptive feedback from an earlier portion of a movement sequence to cue the timing of a later portion, and considerable evidence was found to support this point of view. The Adams and Creamer decay hypothesis states that the subject uses the decay of proprioceptive feedback in short-term memory to cue the timing of a latter movement, and only recently has clear support for this view been provided. Finally, there was the possibility that feedback is not used at all, with the timing of the various sequences being programmed.

The high levels of skill displayed by well-practiced subjects are highly dependent upon the timing of the various movements, that is, unless the submovements occur at the proper time, the total response appears jerky and "uncoordinated." Adams (1966) concluded that timing was one of the most important and least studied aspects of human motor behavior, and, in fact, Michon (1967) stated that it is "the fine structure of the timing between successive elements in a string of actions which characterizes skilled performance [p. 1]." As important as timing seems for the understanding of the mechanisms of motor skills learning and performance, the earliest mention of the role of timing in skill came as late as 1949 from Helson, who demonstrated that subjects in a tracking task without preview were able to track a predictable sine-wave target course with lags that were very near zero. Helson (1949) concluded that the subjects were learning the regularities of the target course and were anticipating its changes so that timing errors were very small. This ability of subjects to anticipate is extremely important for the understanding of motor skills, for without the ability to anticipate the arrival of future stimuli (e.g., a rolling ball), the subject would be forced to respond with drastic delays due to reaction time. Thus, being able to anticipate allows the subject to ready himself for the upcoming event and to respond with the near zero lags that Helson noted.

Timing has received most of its study in a motor skills context using tasks in which the subject tries to anticipate the arrival of some event in time and is required to make some motor response such that the response occurs with a predefined relationship to the external event. An example is hitting a ball, in which the subject anticipates the arrival of the ball and times his response so that the ball is hit when it is in a certain location. In the context of making responses generally coincident with some external event, Poulton (1952, 1957) has defined two classes of anticipation. *Receptor anticipation* is involved when the subject can preview the arrival of the upcoming stimuli to which the response is to be made. The word "receptor" implies that the subject is using his various extero-receptors (e.g., eyes, ears) to preview the events. This class of timing is of very great concern to motor skills since so many tasks (e.g., hitting a pitched ball) require preview and a timed

---

[1] Appreciation is extended to Jack A. Adams, Robert M. Christina, Michael J. Ellis, Judith L. Smith, and George E. Stelmach for the helpful comments on an earlier draft of this paper. Requests for reprints should be sent to Richard A. Schmidt, 5 Barbour Gymnasium, University of Michigan, Ann Arbor, Michigan 48104.

© 1971 by the American Psychological Association, Inc.

response. *Perceptual anticipation,* on the other hand, is concerned with situations in which the subject cannot directly preview the upcoming stimulus events, but in which the events occur in some pattern of regularity that the subject can learn. Thus the subject has a learned basis for anticipation, and, once the pattern is learned, he can respond without the use of exteroceptors, the entire response being "internalized," and under the control of internal mechanisms. Michon (1967) has outlined a number of hypothesized internal mechanisms for timekeeping behavior, such as a purely "cognitive" hypothesis in which the subject merely counts and responds when the count has reached a certain total, a hypothesis depending upon scanning electroencephalogram waves, or a hypothesis dealing with neural time "quanta." However, Michon (1967) and Ellis (1969, 1968 [2]) have indicated that the evidence for such hypotheses has generally been negative, and a discussion of this evidence in detail would be beyond the scope of this study. On the other hand, Adams and Xhignesse (1960) pointed out that it is in keeping with current ideas in physiology to have timekeeping be noncognitive and dependent upon neural proprioceptive networks (Ruch, 1951). Thus a logical choice for the mechanism of the internal control for timing is proprioception, since it is well known that proprioception is a rich source of stimuli, and it is thought to have importance for the control of motor responses. While Adams (1968) has reviewed the literature and found little support for the idea that proprioception has anything to do with *learning*, and Laszlo (1968) found that proprioception was not necessary for the *performance* of certain motor responses, this does not preclude proprioception from being important in the *timing* of motor responses. The possible relationship between the response-produced proprioceptive feedback and the timing of that response or later responses is the topic of this study. There are many other aspects of timing and anticipation that could be discussed, but for these the reader should consult Michon (1967), Noble and Trumbo (1967), and Schmidt (1968).

PROPRIOCEPTION DEFINED

Historically, "kinesthesis" and "proprioception" have had similar meanings, although technically there is a difference between them. According to Adams (1968), the term kinesthesis was first used by Bastian (1883) to describe the sensation of movement and did not include mechanisms of the inner ear that are important in telling the subject whether or not he is accelerating, rotating, or upside down. Also, kinesthesis is not applicable to the perception of the *position* of the various limbs with respect to each other, as it is only concerned with *movement* of the limbs.

Sherrington (1906) used the term proprioception to mean the sense of movement included under kinesthesis, along with the forces, pressures, tensions, relative positions of the body segments, and the body's orientation and movement in space, and this is the way the term is used here. Kinesthesis is a more limited term, and Adams (1968) pointed out that it would seem desirable to keep it this way. The primary receptors for proprioception are the *proprioceptors,* which lie in the muscles, tendons, and joints and determine movements and positions of the body segments. Sherrington also included the labyrinthian mechanism as a proprioceptor. This study is primarily concerned with proprioceptive feedback as it relates to timing behavior, with specific emphasis upon the role of the proprioceptors and exteroceptors (e.g., eyes, ears).

PROPRIOCEPTION AND TIMING

Probably the first indication that proprioception was related to timing came from Goldstone, Boardman, and Lhamon (1958). They had subjects (young children, college students, and older adults) estimate 30 seconds under varying conditions. Counting aloud, which resulted in greater proprioceptive feedback from muscles of the mouth and larynx, produced greater accuracy than estimating silently. Also, when subjects were asked to remain silent, many of them violated instructions prohibiting other movements (e.g., tapping of the feet, fingers) and created movements of their own, which presumably aided the estimation process. The authors concluded that the timing task of counting was dependent upon proprioceptive feedback cues, and that the proprioceptive feedback cues must be associated with the time intervals in some prior learning situation in order that they be used effectively. Children below the age of 8 years apparently did not have these associations at hand, as they demonstrated markedly inferior estimations. Also, there are a number of common examples that seem to indicate that proprioceptive feedback is related to timing, such as the musician who taps his foot or nods his head (thereby generating proprioceptive feedback) to keep time to the music. Such observations, plus the experimental work of Goldstone, Boardman, and Lhamon (1958), have led to a number of formulations of how proprioceptive feedback and timing might be related, and each of these is discussed in turn, giving the various hypotheses, predictions, and supportive evidence.

---

[2] Ellis, M. J. Possible mechanisms for the human clock. Unpublished manuscript, University of Illinois, Urbana, 1968.

Input Hypothesis

The input view is the most obvious and common-sense explanation for the linking of proprioceptive feedback and timing. This view maintains that if a subject is required to make a series of timed responses, he uses the proprioceptive feedback from some earlier portion of the sequence to trigger a later portion of the response. The term "input" is used because the subject is viewed as using the pattern of input or inflow of proprioceptive feedback from the earlier portions of the response to cue the later portion. Stated this way, the input view is really closely related to stimulus-response theory and the "chaining hypothesis." The chaining hypothesis states that the proprioceptive feedback contains the stimuli to which the responses become *learned*, while the input view goes a bit further and maintains that the proprioceptive feedback contains the stimuli to which the response becomes *timed;* thus, one has to do with *what* occurs and the other with *when* it occurs (Adams, 1968). An example that might illustrate the use of such a mechanism is involved in rhythmic tasks, such as pedaling a bicycle. In this task, the muscular forces for extension of the knee must be applied at a point soon after the pedal has passed the top of the circle, and thus require that the subject "time" the application of force. Eliciting the "command" to extend the knee, neural transmission, and tension development in the muscle take a finite and, under certain conditions (e.g., Santa Maria, 1970; Schmidt & Stull, 1970), substantial amount of time. These delays should not be confused with classical reaction time, since reaction time is the delay in response to stimuli that are not psychologically predictable and are suddenly presented, which is not the case for the present example, since pedaling involves anticipation. Thus the subject cannot use the proprioceptive feedback associated with the leg being in the desired position (i.e., just after the top of the circle) as a cue for knee extension, since the delays in the system would cause the response (contraction of the knee extensors) to occur too late. Thus, it is possible that the subject uses the proprioceptive feedback from an earlier portion of the response (e.g., related to some position on the upswing) to cue the knee extension and that through learning, the subject "chooses" proprioceptive feedback cues that will account for the delays inherent in the system and insure that the response occurs at the correct time. Thus, the input view says that the subject uses the response-produced proprioceptive feedback from earlier portions of a response series as cues for initiating later responses and that the chief advantage of doing so is the elimination of the drastic delays due to the lags inherent in the motor system. Thus, it is possible that the subject has some mechanism that receives the proprioceptive feedback input and cues the desired response. Exactly how this proprioceptive feedback mechanism might operate is certainly open to question, but one might envision the proprioceptive feedback input as having some time-predictable pattern (as a function of the time-predictable pattern of the movements producing the proprioceptive feedback) and that when some stage in this pattern was reached, the mechanism would trigger the desired response automatically. Learning to anticipate, under this view, would involve the strengthening of bonds between the characteristics of the proprioceptive feedback input from earlier movements and the later timed responses in the presence of knowledge of results. Thus, the subject is seen as adjusting the cuing mechanism by selecting as a cue that portion of the proprioceptive feedback input that results in the most accurate timing of the automatically cued response.

A number of testable predictions can be made from this hypothesis. First, since the mechanism is seen as responding to changes in proprioceptive feedback, with some minimum change being required in order that a change be detected, it follows that the greater the rate of change in proprioceptive feedback inflow, the smaller should be the temporal lag in triggering the mechanism. Hence, if the response is to be cued with this mechanism, greater temporal accuracy should result if the rate of change in proprioceptive feedback inflow is greater, and, of course, the lower limit should be the situation in which proprioceptive feedback is not changing at all, which would provide no cuing function. One way to increase the rate of change in proprioceptive feedback inflow, provided that the pattern of proprioceptive feedback inflow is not constant over time, would be to increase the magnitude of the proprioceptive feedback discharge by altering the force and amplitude requirements of the movement producing the proprioceptive feedback (Bahrick, 1957). Thus, the first prediction is that increased proprioceptive feedback from movements early in a sequence should increase the timing accuracy in some later response.

A second prediction is that the proprioceptive feedback must be consistent from trial to trial in order that the subject learn the pattern of proprioceptive feedback input and be able to respond to the proper portion of the stream of proprioceptive feedback input. If the proprioceptive feedback input was variable from trial to trial, the subject would have no time-reliable cue to pair with the timed response, and he would have no means by which he could bring the timed response under proprioceptive feedback control. This leads to the prediction that if the movements in an earlier subresponse in a sequence are different from trial to trial (i.e., inconsistent), the subject will not be able to use heightened proprioceptive feedback cues to aid timing, and heightened proprioceptive feedback cues should have diminished effects on the timing of the later response.

Evidence for the Input Hypothesis

There are a number of studies that indirectly link increases in proprioceptive feedback to improved timing behavior, generally supporting an input view. As mentioned earlier, Goldstone et al. (1958) found that counting aloud (presumably generating greater proprioceptive feedback from the lips, tongue, and larynx) resulted in greater accuracy that counting silently in counting to 30 seconds at the rate of one count per second. Adams and Creamer (1962b) used a manual step-tracking task in which the subject had to keep a pointer between two lines with no preview. At regular intervals of either 2 or 4 seconds, the lines would suddenly disappear and reappear on the opposite side of the display, and the subjects' task was to anticipate the change and to move to the new position as quickly as possible. They manipulated proprioceptive feedback by varying the spring tension on the control (the control was spring-loaded to return to the center position between the two targets) and by varying the amplitude of movement between the two target locations. Using "beneficial anticipations" (the proportion of responses on which the subject began his movement to the new target within 133 milliseconds of the termination of the present target) as the criterion score, they found that spring loading improved timing accuracy, but that amplitude failed significance. It is questionable whether the amplitude manipulation was a fair test for the input view, since the input from amplitude was manipulated while the subject was moving to the new position, and, once arriving, the subject remained motionless until the next stimulus change. Thus, when timing accuracy was measured (the time the subject began his next movement), there was no proprioceptive feedback from arm movement with which the subject could cue the next response. However, the spring tension effects were large and supported the input view. But increasing the amplitude between target positions on the Adams and Creamer apparatus also should have, according to Hooke's Law, increased the tension required to hold the cursor on target (since the spring loading tended to center the cursor between the targets), and one wonders why the amplitude effects were not significant for this reason.

Possibly, increased amplitude created a mechanically unfavorable system for human responding that worked against the proprioceptive feedback effects. Additionally, minimal force cues might be sufficient to provide a basis for timing, and increasing force beyond some minimal value might produce no additional effect.

Grose (1964,[3] 1967) had subjects either lift a finger from a key, move the hand 2 feet and press a key, or walk 9 feet and kick a barrier at the moment a moving pointer (with preview—receptor anticipation) became aligned with a stationary one. He found that as the size of the motor response was increased there was a progressive decrease in algebraic (directional) error, and a decrease in the within-subject variability (i.e., an increase in consistency). One possibility is that the larger responses generated greater proprioceptive feedback, which gave the subjects a better basis upon which to time the arrival of the moving pointer. This is not the only conclusion that can be drawn, however, as the procedures confounded movement time, movement complexity, and proprioceptive feedback, and additional work is needed to separate these effects. Ellis, Schmidt, and Wade (1968) attempted to control for these variables and also to determine the role of different sources of proprioceptive feedback. They used a task in which the subjects had to estimate 2 seconds by moving a slide along a trackway. The interval to be estimated began when the subject began to move, and the subject attempted to arrive at the end of the trackway so that the entire response took exactly 2 seconds. They manipulated the amplitude (2.5 or 65.0 centimeters) and the loading (essentially 0 or 8.5 pounds) of the movement in a 2 × 2 design. They found that increased amplitude of movement resulted in smaller absolute error later in practice, supporting the idea that the increased proprioceptive feedback from the increased movement provided the subjects with an improved basis for timing, but it should be pointed out that amplitude and velocity cues were confounded with this design. Load effects were not significant, creating findings nearly opposite to those of Adams and Creamer (1962b). In another study to determine the effects of various modalities of feedback, Ellis (1969) used a similar task but manipulated the acceleration cues (by manipulating the inertia of the lever system), velocity cues (by manipulating the viscous resistance by an electromagnetic brake), and tension cues (by manipulating the spring tension on the movement) and compared these conditions with no-resistance control conditions. His results generally supported the conclusion that the nature of the resistance (and, hence, the proprioceptive feedback) to the movement caused differential effects on timing accuracy with greater proprioceptive feedback resulting in greater accuracy, greater response consistency, and fewer trials to learn to a criterion. Further, Ellis had subjects spell two-, three-, and four-letter words during their movements, and he found that timing accuracy increased with increasing word length, the conclusion being that increased word length produced greater proprioceptive feedback. Finally, when the spelling was withdrawn, the no-movement control groups regressed drastically in timing, while the movement conditions showed only slight decreases. Apparently, the proprioceptive feedback from spelling gave the subjects some cues for timing, and withdrawing spelling left them with no basis for timing, but the groups with movement could still rely upon the movement cues for timing in the absence of the spelling cues. It seems probable from Ellis' work that subjects do not rely upon such gross mechanisms as counting in everyday tasks requiring timing, and that proprioceptive feedback is used in some way to improve accuracy.

Although these studies seem to implicate proprioceptive feedback in timing somehow, there is a methodological problem that prevents such a strong conclusion. In each case, the proprioceptive feedback was manipulated by changing the dynamics of the control, and thereby changing the proprioceptive feedback may have also created a mechanically more favorable system for responding. Thus, a better test of the role of proprioceptive feedback in timing would involve manipulations of proprioceptive feedback in one limb and have the response effected in another limb. Schmidt and Christina (1969) recognized this limitation and sought to manipulate proprioceptive feedback in the left arm by having the subjects execute a minimal finger movement or a small or large rotary movement. Beginning the left-arm movement also started a target, and the subject responded with the right hand when the moving target became aligned with a stationary one. No preview of the target was given, and essentially the subject's task was to estimate when the target would arrive at the stationary pointer, the travel time being constant at 1.50 seconds. Using "beneficial anticipations" (the percentage of responses that were within 133 milliseconds of correct) as the criterion score, they found that the small rotary movement of the left arm led to significantly greater timing accuracy than the control (finger response) condition, with the large rotary movement falling between these two extremes. The findings supported the view that increased proprioceptive feedback in the left arm contributed to increased timing accuracy in the right arm, and thus the increases in timing found in earlier studies that manipulated proprioceptive feedback in the effector were probably not entirely attributable to purely mechanical effects of control dynamics manipulation. On the other hand, the findings were not completely clear, since, according to the predictions from the input view, the group with the greatest proprioceptive feedback should have been most accurate, which was not the case. Further analysis of the data revealed that the large left-arm movement was far more inconsistent (within subjects) than the smaller left-arm movement (3.5 times as inconsistent), and that it was possible that the inconsistency of left-arm movement in the larger response provided a poorer basis for timing than the smaller, more consistent movement. This provided indirect support for the input hypothesis, since consistency of the proprioceptive feedback input from trial to trial was hypothesized to be one of the relevant variables.

---

[3] Grose, J. E. Timing control in finger, arm, and whole body movements. Unpublished doctoral dissertation, University of California, Berkeley, 1964.

One simple strategy which the subjects could be utilizing the proprioceptive feedback for timing would be to respond with the right hand when the left arm had reached a certain position. Schmidt and Christina used within-subject correlations (Grose, see Footnote 3; Schmidt, 1969) to measure the degree of relationship between the amount of the left-arm movement (the position of the left arm when the subject responded with the right hand) and the algebraic error (with respect to sign) in responding with the right hand over a series of trials. Thus, each subject had a correlation coefficient calculated for him, the hypothetical N for the coefficient being the number of trials (i.e., 15) over which the coefficient was calculated. Findings indicated that exactly half of the subjects tended to covary the left-arm movement and the time of responding with the right hand, showing correlations greater than the .40 needed for significance at the .01 level.

Christina (1969,[4] 1970) used the same right-arm task as did Schmidt and Christina (1969), but investigated the effects of tension and movement proprioceptive feedback. He had subjects make either a small (.5-inch) hand movement, a 13-inch linear movement of an unloaded slide, or a 13-inch linear movement of a loaded slide, each with the left arm. The apparatus was like that of Schmidt and Christina's (1969) in that the subject's initial movement of the slide started the target, and the subject moved the left hand in one direction until he responded to the main task with the right hand. He found that the groups with left-arm movement performed the right-arm timing task more consistently than the no-movement control group, with the load effects not being significant. However, there were no significant differences between treatments using absolute error, algebraic error, beneficial anticipations, or number of trials to a criterion. The finding of no beneficial anticipation effects was contrary to earlier findings by Schmidt and Christina (1969).

In summary, earlier evidence has indicated that greater proprioceptive feedback by manipulating the control dynamics in the effector improved timing, but these findings confounded mechanical and proprioceptive effects. Thus, the effects of spelling withdrawal by Ellis (1969), and the findings of

---

[4] Christina, R. W. Proprioception as a basis for the anticipatory timing of motor responses. Unpublished doctoral dissertation, University of Maryland, 1969.

Schmidt and Christina (1969) and Christina (1970; see Footnote 4) with left-arm proprioceptive feedback manipulations and a right-arm task, are the only sources of evidence that clearly link proprioceptive feedback to timing performance.

Decay Hypothesis

The proprioceptive feedback decay hypothesis was first stated by Adams and Xhignesse (1960) to account for the timing of responses when there was no predictably changing source of proprioceptive feedback input (e.g., with perceptual anticipation). Essentially the same ideas were presented again by Adams and Creamer (1962a, 1962b) and were formally tested by Adams and Creamer (1962b). The hypothesis states that if the subject makes a response at time *t*, the proprioceptive feedback from the response enters a short-term storage system (possibly short-term memory) and decays predictably as a function of time. The pattern of decay is not specified, but the assumption is that a given movement will produce a predictable input of proprioceptive activity, and this activity will decline predictably as a function of time in short-term memory. Now, if an additional response is to be made at time $t + \Delta t$ (Δt a constant), the decay view states that the subject uses the time-varying properties of the proprioceptive trace in short-term memory as a time standard, possibly responding when the proprioceptive feedback activity had decayed to a given level, thereby providing the subject with a basis for cuing the later timed response.

A common example in which the proprioceptive feedback decay hypothesis might apply is springboard diving. The proprioceptive feedback cues associated with acceleration off the board would enter short-term memory and begin to decay, and the diver might use the decaying proprioceptive feedback cues as a basis upon which to make some later response, such as a precisely timed "untucking" while he is in the air. The decay hypothesis is thus distinct from the input view stated earlier, since the decay view states that the diver uses as a cue the decay of proprioceptive feedback from an earlier movement (e.g., acceleration from the board), while the input view states that the diver uses proprioceptive feedback input from some concurrent movement (e.g., the proprioceptive feedback cues from rotation while he is in the air).

There are a number of testable predictions that come from the basic hypothesis. As with the input mechanism discussed earlier, the decay mechanism is seen as a change detector with some minimum change required to activate it. Thus, the greater rate of change in proprioceptive feedback decay, the smaller will be the lag in activating the mechanism, and a response triggered by the mechanism should be executed more accurately also. One way to increase the rate of proprioceptive feedback decay, provided that the pattern of decay is exponential or at least not linear (e.g., a decreasing negatively accelerated function of time), would be to increase the absolute amount of proprioceptive feedback activity placed in short-term memory, which should result in a greater decay of proprioceptive feedback in the first few seconds. This leads to the prediction that if the preliminary movement at time *t* has large force or amplitude requirements so that it generates considerable proprioceptive feedback, then there should be greater proprioceptive feedback in short-term memory, a greater rate of decay in the next few seconds, and less lag in the mechanism for triggering the subsequent response, with the result that the timing of the response at time $t + \Delta t$ will be more accurate.

Second, if the pattern of decay is decreasing and negatively accelerated, the rate of decay will be decreasing with each successive moment in time. If the interval Δt is short, the mechanism will be triggered when the rate of decay is great, and when the interval Δt is long, the mechanism will be triggered when the rate of decay is relatively smaller. Thus, since the lag in the mechanism is dependent upon the rate of decay, the mechanism should be less effective at long intervals and more effective with very short intervals. This leads to the prediction that the decay mechanism should be very effective, relative to a no-proprioceptive-feedback control condition, when the interval is short, but should be no more effective than a no-proprioceptive-feedback control condition when the interval is very long. In other words, there should be an interaction between presence or absence of proprioceptive feedback from a prior movement and the length of the interval (Δt) to be estimated.

Third, if the preliminary movement were inconsistent from trial to trial as the subject was learning the task, he would have a different amount of proprioceptive feedback introduced into short-term memory on each trial, and the activity would decay to the cuing level in a different number of milliseconds for each trial. If this be the case, the subject would have no basis for pairing the decay of feedback with the time interval to be estimated (the time from t to $t + \Delta t$), and performance would not be better than if some "cognitive" mechanism were used. Thus, the timing of the second response should be a function of the consistency (of force, amplitude, direction, etc.) of the first response, with the greatest timing accuracy being produced when the first response is as consistent as possible from trial to trial.

Evidence for the Decay Hypothesis

Adams and Creamer (1962b) claimed support for their hypothesis using a step-tracking task with a manual lever control. The subject had to keep his lever positioned between two lines at one side of the display, and, when the lines suddenly stopped and reappeared at the other side of the display, he had to move quickly so that he was on target again. The subject could not see directly when the lines would be discontinued (no preview), but the changes were regular at every 2 or 4 seconds, and the subject could learn this and predict the change of position in advance (perceptual anticipation). Adams and Creamer manipulated the force necessary to hold the lever on target (the control being spring-loaded to return to the center of the display) and the movement amplitude. They predicted that the proprioceptive feedback generated by response n in a series would be used by the subject to time response $n + 1$, and that greater proprioceptive feedback from response n would result in increased timing accuracy for response $n + 1$. They found that spring loading caused a significant increase in timing accuracy when they used "beneficial anticipations" (the number of responses on which the subject started his movement to the new position within 133 milliseconds of when the lines terminated) as the score, but amplitude effects were extremely small and nonsignificant. The decay hypothesis predicts that shorter intervals should be more accurately estimated than longer ones, which Adams and Creamer found, but this finding does not differentiate between the decay or purely "cognitive" (e.g., counting) hypotheses, which both predict poorer timing with longer intervals.

Although the Adams and Creamer decay hypothesis has been generally accepted as an empirically supported mechanism for timing (e.g., Singer, 1968, p. 74), there are some methodological problems with the Adams and Creamer experiment that warrant a more cautious approach. First, in order to demonstrate that it was the decay of the proprioceptive feedback which the subject was using, it is necessary to have the subject be at rest during the interval from the first movement until the second movement; otherwise, one could not be sure that the better performance as a function of the control dynamics was not due to the input of proprioceptive feedback during the interval (the input hypothesis). However, Adams and Creamer's subjects had to hold the lever in place against spring tension during the interval to be timed, and it could be that the spring tension effects were due to the subjects using the proprioceptive feedback *input* from static tension during the interval. Thus, although the spring tension effects were significant, they do not differentiate between the Adams and Creamer decay hypothesis and the input hypothesis. The amplitude dimension of the Adams and Creamer experiment gave a valid test of the decay hypothesis since the limb was motionless during the interval to be timed, but the amplitude effects were not significant. Second, the proprioceptive feedback from spring loading was generated in the same limb that effected the response, and one cannot be sure that the spring system did not simply create a mechanically more favorable system for responding. Thus, a valid test of the Adams and Creamer hypothesis would involve a situation in which the subject was motionless (with no static tension) during the interval to be timed, and in which the proprioceptive feedback was manipulated by manipulating control dynamics in a different limb from the one which is the effector for the timed response.

Christina (1970; see Footnote 4) recognized the design deficiencies of the Adams and Creamer experiment and gave the decay hypothesis an additional test. He used a discrete task in which the subject had to make a finger response to estimate the end of a 2-second interval and manipulated proprioceptive feedback in a prior movement by having no movement, unloaded movement, or loaded movement, each performed prior to the interval to be timed. His subjects (in the movement conditions) made a linear 13-inch movement of a slide on a trackway with the left hand. When the subject reached the end of the trackway, a switch started the interval to be timed, and the subject responded 2 seconds later with a right-hand response, the left hand remaining motionless during the interval. If the decay hypothesis is true, subjects with prior movement would be superior to the control, and prior movement plus load would be superior to the movement alone. Results were negative, with no significant differences shown for algebraic error, absolute error, and "beneficial anticipations," which did not support the decay view.

However, the proprioceptive-feedback-generating left-arm movements in the Christina (1970) experiment were not perfectly consistent from trial to trial, and since consistency seems necessary for the decay method to be effective, this may have prevented the decay from providing an effective basis for timing. Christina did attempt to have his subjects execute the left-arm responses as consistently as possible, providing preliminary practice at the left-arm task alone prior to actual testing to emphasize consistency, and, during the right-hand practice, providing knowledge of results to inform the subject whether he was responding within the acceptable a priori limits for the left-hand movement. Even so, the left-hand movements showed considerable variability within subjects over trials. Thus, Quesada and Schmidt (1970) attempted to make the left-arm responses more consistent by using a passive, motor-driven movement. Subjects in a movement group had their left arm moved passively through 26 inches at a constant speed so that the movement time was nearly consistent at 1.42 seconds on every trial. At the end of the movement, the subject's arm came to rest on the arm of a chair. Control subjects had complete view of the moving handle, but did not hold onto it, their arms resting on the arm of the chair during the handle movement. The end of the handle movement signaled the beginning of the 2-second interval to be timed (also signaled by a light and a buzzer), and the task was to lift the right index finger from the key when 2 seconds had elapsed. On Day 1, subjects had 40 trials with knowledge of results, and the movement subjects demonstrated improved timing accuracy with smaller within-subject variability and absolute and algebraic error, which supported the decay hypothesis. On Day 2, one week later, subjects were given 10 additional trials with knowledge of results, followed by 20 trials on which knowledge of results was withheld. It has been found (Ellis et al., 1968) that knowledge of results withdrawal leads to a drift in the direction of overestimating time, and support for the decay hypothesis would be provided if there were less drift for the movement group. This was found to be so, with considerably less absolute error and within-subject variability for the movement group near the end of the 20-trial knowledge-of-results withdrawal period. Thus, these two findings provided two lines of support for the decay hypothesis, and indicated that Christina (1970) may have failed to provide support in his experiment because of the inconsistency of the left-arm movements.

The Quesada and Schmidt study was the only one found that supported the decay hypothesis, and so has provided a basis from which to investigate further the generality of the decay view. Research is currently under way to investigate some of the other predictions of the hypothesis, such as the effects of left-arm consistency and of the length of the interval (t to $t + \Delta t$) to be estimated, to determine whether this hypothesis is appropriate for other classes of proprioceptive feedback (e.g., force, vestibular), and to determine whether the hypothesis is supported for tasks using receptor anticipation. It is certainly an interesting hypothesis, one which could go a long way to explain the nature of the cuing process for the timing of motor responses.

Preprogramming Hypothesis

Of course, it is possible that either or both of the proprioceptive feedback hypotheses are wrong, and that the subject responds on the basis of a neuromotor program without relying upon proprioceptive feedback. Under this view, the subject has a program that dictates the timing of the various submovements in a sequence, and once the program is triggered, the timing of the sequence is smoothly and precisely executed. This idea has been presented by Fitts (1964), who referred to a hierarchical system of motor control implying that there are higher centers to select the correct program and lower centers that contain the program and execute it upon command from the higher center. A similar concept has been discussed by Henry and Rogers (1960). It is beyond the scope of this study to discuss the evidence for program versus feedback control of motor responses, and for reviews of this area see studies by Adams (1968) and Keele (1968); but one study in this area by Pew (1966) should be mentioned, since it has special relevance to timing behavior. He used a two-hand task in which the subject attempted to keep a dot centered on a display. Pressing a button with the right hand made the dot accelerate to the right, and pressing a button with the left hand made it accelerate to the left. Pew's records indicated that later in practice the subjects were responding with very rapid alternating button-pressing responses, which were much too rapid to be explained by a mechanism in which the subject pressed the button, waited for the visual consequences of his act, and then pressed the button again to make a correction. It was as if the subject was pressing the buttons under the control of a central program, and occasionally the subject would modify the program by changing the subtle timing relationship between the two hands. For example, with the subject responding very rapidly, if the target began to creep to the right, the subject would change his timing so that the left-hand button was held down slightly longer than the right button, but would continue to press alternately at the fast rate. This evidence indicates that the timing of the responses was not under feedback control, since subjects were able to make only two to three corrections per second, which was much slower than Pew's subjects, and might indicate that the timing of the responses was truly preprogrammed. However, since Henry (1953) has shown that subjects can respond to proprioceptive feedback without perceiving it (also Posner, 1969), it is certainly possible that Pew's subjects were responding with a proprioceptive feedback basis, with the closed-loop control being sufficiently fast to account for the rapid rate of button pressing. Also, a preprogramming view that predicts that subjects time without any reliance on proprioceptive feedback would have trouble accepting the findings of Ellis et al. (1968), Ellis (1969), Schmidt and Christina (1969), Christina (1970), and Quesada and Schmidt (1970), who all found that enhanced proprioceptive feedback contributed to improved anticipation. It is certainly possible that timing is governed by programs only in certain situations, and determining the laws governing when timing is preprogrammed and when it is proprioceptive feedback controlled will be a major research task.

DISCUSSION

The evidence presented here suggests strongly that proprioceptive feedback and timing are related, and hypotheses were presented that predict different ways that the proprioceptive feedback might be utilized to provide a basis for timing of motor responses. There is a great deal of support for an input hypothesis, which states that the subject uses the pattern of proprioceptive feedback inflow from earlier movements in a movement sequence as a basis for timing, the subject executing the timed response when the pattern of input has reached some state that has been learned. After methodological problems were eliminated, empirical support was found for the Adams and Creamer decay hypothesis in which the subject used the decay of proprioceptive feedback in short-term memory as a basis for timing, but it should be tested further in a number of ways in order to determine its generality. Finally, there was the possibility that proprioceptive feedback is not involved in timing at all, with timing of the various movements in a sequence being governed by motor programs that spell out the precise temporal relationships among the responses in the sequence. Of course, it is possible that all three mechanisms are available to subjects, and that different mechanisms are used for tasks with different display and response requirements. Determining if, or under what conditions, each of these theoretical propositions is appropriate for the explanation of timing in motor behavior will be a major undertaking, but a thorough understanding of motor behavior will await the answers.

REFERENCES

Adams, J. A. Some mechanisms of motor responding: An examination of attention. In E. A. Bilodeau (Ed.), *Acquisition of skill.* New York: Academic Press, 1966.

Adams, J. A. Response feedback and learning. *Psychological Bulletin,* 1968, **70**, 486–504.

Adams, J. A., & Creamer, L. R. Anticipatory timing of continuous and discrete responses. *Journal of Experimental Psychology,* 1962, **63**, 84–90. (a)

Adams, J. A., & Creamer, L. R. Proprioception variables as determiners of anticipatory timing behavior. *Human Factors,* 1962, **4**, 217–222. (b)

Adams, J. A., & Xhignesse, L. V. Some determinants of two-dimensional visual tracking behavior. *Journal of Experimental Psychology,* 1960, **60**, 391–403.

Bahrick, H. P. An analysis of stimulus variables influencing the proprioceptive control of movement. *Psychological Review,* 1957, **64**, 324–328.

Bastian, H. C. *The brain as an organ of the mind.* New York: Appleton, 1883.

Christina, R. W. Proprioception as a basis for the temporal anticipation of motor responses. *Journal of Motor Behavior,* 1970, **2**, 125–133.

Ellis, M. J. Control dynamics and timing a discrete motor response. *Journal of Motor Behavior,* 1969, **1**, 119–134.

Ellis, M. J., Schmidt, R. A., & Wade, M. G. Proprioception variables as determinants of lapsed-time estimation. *Ergonomics,* 1968, **11**, 577–586.

Fitts, P. M. Perceptual-motor skill learning. In A. W. Melton (Ed.), *Categories of human learning.* New York: Academic Press, 1964.

Goldstone, S., Boardman, W. K., & Lhamon, W. T. Kinesthetic cues in the development of time concepts. *Journal of Genetic Psychology,* 1958, **93**, 185–190.

Grose, J. E. Timing control in finger, arm, and whole body movements. *Research Quarterly,* 1967, **38**, 10–21.

Helson, H. Design of equipment and optimal human operation. *American Journal of Psychology,* 1949, **42**, 473–479.

Henry, F. M. Dynamic kinesthetic perception and adjustment. *Research Quarterly,* 1953, **24**, 176–187.

Henry, F. M., & Rogers, D. E. Increased response latency for complicated movements and a memory drum theory of neuromotor reaction. *Research Quarterly,* 1960, **31**, 448–458.

Keele, S. W. Movement control in skilled motor performance. *Psychological Bulletin,* 1968, **70**, 387–403.

Laszlo, J. I. The role of visual and kinaesthetic cues in learning a novel skill. *Australian Journal of Psychology,* 1968, **20**, 191–196.

Michon, J. A. *Timing in temporal tracking.* Soesterberg, The Netherlands: Institute for Perception RVO-TNO, 1967.

Noble, M., & Trumbo, D. The organization of skilled response. *Organizational Behavior and Human Performance,* 1967, **2**, 1–25.

Pew, R. W. Acquisition of hierarchical control over the temporal organization of a skill. *Journal of Experimental Psychology,* 1966, **71**, 764–771.

Posner, M. I. Reduced attention and the performance of "automated" movements. *Journal of Motor Behavior,* 1969, **1**, 245–258.

Poulton, E. C. Perceptual anticipation in tracking with two-pointer and one-pointer displays. *British Journal of Psychology,* 1952, **43**, 222–229.

Poulton, E. C. On prediction in skilled movements. *Psychological Bulletin,* 1957, **54**, 467–478.

Quesada, D. C., & Schmidt, R. A. A test of the Adams-Creamer decay hypothesis for motor response timing. *Journal of Motor Behavior,* 1970, **2**, 273–283.

Ruch, T. C. Motor systems. *Handbook of experimental psychology.* New York: Wiley, 1951.

Santa Maria, D. L. Pre-motor and motor reaction time differences associated with stretching of the hamstring muscles. *Journal of Motor Behavior,* 1970, **2**, 163–174.

Schmidt, R. A. Anticipation and timing in human motor performance. *Psychological Bulletin,* 1968, **70**, 631–646.

Schmidt, R. A. Movement time as a determiner of timing accuracy. *Journal of Experimental Psychology,* 1969, **79**, 43–47.

Schmidt, R. A., & Christina, R. W. Proprioception as a mediator in the timing of motor responses. *Journal of Experimental Psychology,* 1969, **81**, 303–307.

Schmidt, R. A., & Stull, G. A. Pre-motor and motor reaction time as a function of preliminary muscular tension. *Journal of Motor Behavior,* 1970, **2**, 96–110.

Sherrington, C. S. *The integrative action of the nervous system.* New York: Scribner, 1906.

Singer, R. N. *Motor learning and human performance.* New York: Macmillan, 1968.

(Received May 4, 1970)

Psychological Bulletin
1971, Vol. 76, No. 6, 394–408

EFFECTS OF MEMORY TIME ON SUCCESSIVE JUDGMENTS

J. D. TATE [1] AND R. M. SPRINGER [2]

California State College at Los Angeles

The apparent changing of stimulus values during intervals between successive comparisons is both a practical problem in real life and a problem for research employing psychophysical procedures. The literature of two distinct sources is reviewed: color memory and time error studies. Contradictory results are found both between and within these sources, though there has been a tendency for the former to find predominantly positive time error and the latter, negative time error. Contextual effects are apparently very important. Discrepant results are also attributed to such factors as differences in stimuli and psychophysical methods employed, methods of data analysis, peripheral aftereffects, adaptation, and practice. A critique is offered and the results of a "model" study are cited that indicate that time errors are seriously affected by the inclusion of poor discriminators and that sex differences in color memory are important.

The change in judgments of physical stimuli as a function of the time between the presentation of a "standard" stimulus and a "comparison" stimulus is a problem that has plagued psychologists since the time that Fechner introduced his psychophysical procedures. The changes that take place have both practical and theoretical importance. In psychophysical procedures they must be controlled in order that the judgments will reflect the subject's sensitivity as closely as possible. Often, however, these systematic errors cannot be balanced, and we need to be able to describe the function relating judgments to the time interval between stimuli and explain why these changes take place. In real life it is frequently necessary to make successive judgments with a time lapse, and, as a matter of fact, the present study had its origin in one such practical problem. During the fabrication of tiny, integrated electronic-circuit components, inspectors are required to remember, with a high degree of accuracy, a greatly magnified color patch and match it to a set of color standards.

Two distinct sources were consulted to determine the matching errors that occur during successive comparisons:

1. Color memory research. Although not conclusive, relatively recent studies on color memory by optical physicists indicated a large apparent intensification of color in the memory, especially for the saturation and brightness aspects.

2. Time error research. During the 1930s, considerable effort was expended by psychologists to determine the errors that occur during successive matching of stimuli in any sensory modality; the research was known as the study of the "time error." Errors of apparent stimulus intensification in memory were known as positive time errors; an apparent fading in memory was a negative time error. The time error studies clearly indicated a negative time error for the intensive dimensions of most stimuli, including the saturation and brightness aspects of color. The time error and color memory research results are thus clearly in conflict.

The purpose of the present study is to (*a*) review the literature relevant to the direction and magnitude of errors that occur during successive judgments; (*b*) discuss factors underlying the conflicting results; and (*c*) summarize and evaluate the adequacy of the studies.

COLOR MEMORY RESEARCH

The study of memory for color can be traced to Hering (Boring, 1942), although some research was performed prior to Hering (Bentley, 1899). Hering felt that color memory played a major role in color constancy. Katz (1935) deemphasized Hering's role for color memory by pointing out the stability of the colors of unfamiliar objects. Following Katz, color memory research waned until 1942, when Hanawalt and Post performed an intensive four-part study to investigate saturation memory.

---

[1] Requests for reprints should be sent to J. D. Tate, Department of Psychology, California State College, 5151 State College Drive, Los Angeles, California 90032.

[2] Now at the University of California, San Diego.

During the first part of their study, subjects were shown color chips, which they attempted to match from nine very similar alternatives. The experimental conditions were either a delay of 1 second or 1 week between the presentation of the standard stimulus and the comparison stimulus. For the first condition they found a positive time error, but no systematic error was found for the 1-week interval.

In the second part of the study, a method of limits, using a color wheel, was employed. Negative time errors were found on "ascending" series (an increasing proportion of the matching color) and positive time errors for the descending series. An average of ascending and descending series gives a nearly zero time error.

Saturation memory was the subject of the third and fourth parts of the Hanawalt and Post study. In the third part, subjects were shown standards of either 50% or 75% saturation and were asked to indicate whether a comparison stimulus identical to the standard was more or less saturated than the standard. They did not find time errors for either standard at either the 28- or 58-second interval. In the fourth study part, standards of 20%, 75%, and 95% saturation were used. After 48 hours, the subjects were asked to determine if an identical comparison stimulus was of equal, more, or less saturation than the standard. They found negative time errors for the 20% and 95% levels, but no error for the 75% standard.

Hamui and Landis (1955) had subjects match each of 10 Ostwald chips to a set of 672 alternatives after memory spans of 15 minutes, 24 hours, and 65 hours. The authors made no conclusions for hue, but they found negative time errors for brightness and positive time errors for saturation. The effect was independent of memory time.

An intensive and carefully controlled study of color memory was undertaken by Newhall, Burnham, and Clark (1957), who were perplexed by the lack of knowledge and agreement on errors in color matching and the great need for such knowledge: "whenever a color inspector or a technical color tester compares a color sample in one location to a color standard in another location, color memory is necessarily involved [Burnham & Clark, 1955, p. 164]." Three subjects were shown 25 standards on a colorimeter screen. The 25 standards were samples of different hues, brightnesses, and saturations. After a 5-second memory interval with a blank screen, each subject readjusted the colorimeter until a match to the standard had been achieved. Match times averaged 25 seconds after the memory interval ended. The authors found no constant error for hue; however, they obtained a large positive time error for saturation (1.7 Munsell chroma points) and brightness (.18 Munsell value points). Moreover, they found the size of the time error was highly correlated with the level of the standard. The authors advanced the hypothesis that the saturation and brightness of a color is intensified in memory in proportion to the absolute intensity level of the stimulus.

Three supplementary studies are also reported by Newhall et al. (1957) as providing relevant data on this topic: (*a*) A study by Burnham is essentially a replication of the preceding experiment; the results are almost identical, though match times average about 50 rather than 25 seconds. (*b*) Burnham and Clark had subjects match Munsell chips from nine similar alternatives after a 10-second memory interval. They found, in agreement with the two previous studies, a positive time error for saturation that increases with more intense standards. However, they found no time error for brightness. (*c*) Newhall and Pugh asked subjects to select Munsell chips that matched a variety of absent familiar stimuli for which the average color values were known. They found hues shifted toward verbal norms; in addition, they found positive time errors for saturation and brightness.

Bartleson (1960) ran an experiment similar to that of Newhall and Pugh. He too found that hue shifted toward verbal norms. Positive time errors were obtained for saturation and brightness—for instance, red bricks got redder, grass greener.

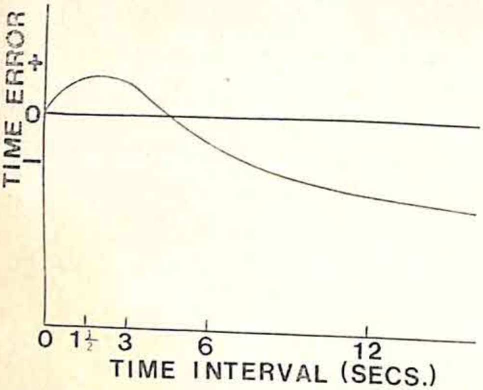


FIG. 1. Köhler's *p*-function for loudness of click stimuli.

Summary of Color Memory Research

The early research by Hanawalt and Post (1942) and Hamui and Landis (1955) was equally divided among findings of positive, negative, and negligible time errors for saturation and brightness. Little data were obtained for hue. The later research initiated by Newhall et al. (1957), however, is in complete agreement that for saturation and brightness positive time errors occur (average 1.0 Munsell chroma steps for saturation and .5 value steps for brightness). Some evidence was also found to indicate that hues move toward verbal norms.

TIME ERROR RESEARCH

In general, the time error studies did not test memory for color; however, despite Christal's (1958) factor-analytic finding that color memory is a unique memory factor, the time error studies are of great importance to an understanding of errors in color memory also. If similar time error functions are found during successive comparisons for all stimuli other than color, then it should be expected that similar error functions will exist in color memory.

Fechner's Memory Image Theory

Much of the earlier time error theory and research are summarized by Needham (1934a). He pointed out that it was Fechner who first noticed the time error as a constant subject bias that produced asymmetrical judgments with a symmetrical stimulus series in psychophysical procedures. Fechner formulated an image theory for the constant errors, which he called "time errors"; his explanation can be traced to Aristotle. According to this concept, a subject obtains an image of the standard that fades with time, resulting in an apparent intensification of the comparison stimulus—a negative time error. In order to remove the time error as a source of systematic error in psychophysical methods, Fechner introduced the counterbalancing method in which the comparison stimuli are presented before and after the standard stimulus an equal number of times.

Needham (1934a) pointed out that researchers not only found Fechner's technique inconvenient but it did not entirely eliminate time errors. Moreover, the correction method inhibited research on time error until the 1920s (Pratt, 1933a), with the exception that numerous critiques of the memory image concept were aroused. Postman (1947) cited studies with tones, brightness, and weight in which time errors with imageless comparisons were found.

Köhler's Fading Trace Formulation

Wolfgang Köhler began an investigation of the time error problem in 1923, not to refine psychophysical techniques, but as a source of valuable information about simple sensory mnemonic traces (Marchetti, 1942). In one study, Köhler tested 12 subjects with click stimuli of varying loudness. He found that for memory intervals to 3 seconds, the time error function was positive. At 3 seconds, the function became zero, and beyond this the time error grew progressively more negative. This function is shown in Figure 1.

Köhler called this function relating memory time and time error a *p*-function for pause or interval (Needham, 1934a, 1935a). He hypothesized that a lag exists between the offset of the stimulus and the maximum buildup of the H ions in the brain; the buildup creates the positive portion of the *p*-function (Freeman & Sharp, 1941). Following the peak buildup, a continuous dispersion of ions (dissimilation) ensues. "This phase of activity was a non-process condition, that is, it did not have a conscious correlate [Needham, 1934a, p. 236]." Köhler's research stimulated numerous studies and received much support.

Visual research. Bentley's (1899) early work with brightness revealed no time errors for 2- and 6-second memory spans. Increasingly more negative errors were obtained for intervals over 10 seconds (tested to 5 minutes). Koester (1945) indicated that Fodor and Happisch also obtained a negative time error for brightness with a 30-second interval.

Peak (1939) presented standard and comparison stimuli only 177 and 285 milliseconds apart. In both cases she found positive time errors for brightness, as predicted by Köhler; the 285-millisecond interval had a slightly larger error.

Kreezer (1938) successively presented light either to the same retinal spot or to opposite sides of the retina. Light to opposite sides produced only a negative time error, while light to the same spot yielded a decreasing positive curve. He theorized that the positive portion of the *p*-function for brightness is an artifact of a local "inhibitory aftereffect."

An unpublished study by Crannel, reported by Tresselt (1944b), tested time errors in judgments of the visual extent of successively projected dotted lines on a screen. He found that the dots were remembered as being closer together than their true separation; he interpreted this as a negative time error. The errors increased with longer memory times. In another study of extent, Allport (1931) monitored the successive reproductions of Binet x_{13} designs by 350 children. He found that the reproductions grew smaller, and this was also interpreted as a negative time error.

Kinesthetic research. Payne and Davis (1940) utilized a single subject to obtain data on the successive comparisons of lifted weights. The authors found a *p*-function identical to Köhler's; however, they interpreted the results in terms of fading muscle potentials.

Freeman and Sharp (1941) obtained a *p*-function similar to Köhler's for intervals up to 30 seconds; however, at 60 seconds the error was less negative than at 30 seconds. Although they agreed with Köhler's theory, they also felt that a major factor influencing such results is "differential backlash excitation from the periphery."

Weinstein (1955) performed a study with weights which is analogous to Kreezer's same- or opposite-side retinal study. His subjects lifted weights either alternately with each hand or successively with one hand. The results are also analogous to Kreezer's because there is a more positive time error for the same hand (same area). This again indicates the possible importance of peripheral aftereffects in determining the shape of the *p*-function.

Auditory research. Needham (1934a) replicated Köhler's experiment and confirmed the original findings, with the exception that the transition from positive to negative errors occurred at 2 rather than 3 seconds.

Postman (1946) hypothesized that if time errors are the result of a fading trace, then one should predict that sounds varying in intensity will produce much larger time errors than sound varying in pitch, that the pitch of a faded sound will be identifiable, and that the original loudness level will be altered. The results support this hypothesis—there was a significant time error for loudness, but not for pitch. Tones varying in pitch were accurately matched even after intervals of 20 seconds.

Tresselt (1944b) reported that Danzfuss found memory for the aesthetic value of chords followed Köhler's *p*-function.

Other research. According to Koester (1945), Fodor and Happisch tested time error during successive comparisons of salt taste. They found Köhler's *p*-function for intervals to 40 seconds.

Koffka's Intensification Hypothesis

In accordance with the principles of "pointing" or "emphasis," Koffka (1935) hypothesized that the intensity of stimuli (such as loudness, weight, or saturation) would intensify in memory, and to a greater degree with more intense stimuli, longer memory times, and more familiar objects (Davis, 1952).

Visual research. Warren and Shaw (1895) had subjects successively compare the size of 150-millimeter squares with stimuli differing by a few millimeters. They obtained a large positive time error (apparent size increase in memory), which increased steadily with longer memory intervals (to 20 minutes).

Marchetti (1942) also tested time errors for the visual extent of lines and obtained a function exactly inverse to Köhler's; that is, it begins as negative and becomes increasingly more positive (line remembered as longer than actual) for intervals greater than 2 seconds. This result was confirmed by McClelland (1943). For the memory of color saturation, Collins (1931) found positive time errors.

Kinesthetic research. Newhall (1932) studied the judgment of pressure of cutaneous points stimulated successively and reported a positive time error for intervals to 10 seconds. He cautioned, however, that his results were confounded by fatigue effects from the first stimulation which would tend to produce a positive error.

Auditory research. Needham (1934a) cited a study by Wada in which there is an apparent shift in memory to higher pitches (a result contradictory to Postman, 1946); he interpreted this as a positive time error. Higher pitched standards moved the entire *p*-function to a more positive level.

Findings not Supporting a Positive or Negative Time Error Theory

Visual research. Katz (1935) predicted that stimuli would be exaggerated in memory away from a norm. Longer memory intervals were predicted to increase the error; highly saturated colors (such as red blood), for example, would become more intense; poorly saturated colors (such as cement) would fade. Katz presented evidence to support his hypothesis.

Irwin and Seidenfeld (1937) successively presented different types of visual stimuli such as dotted and solid lines, circles, and triangles. Subjects were asked to remember different aspects of the standard such as area or size of edge. They found either positive or negative time errors, depending upon the particular figure presented and aspect judged; no increases or decreases in time errors were observed with changes in the memory interval (3 minutes, 10 minutes, 7 days).

Lesser (1959), in an experiment similar to Irwin and Seidenfeld, had subjects successively match broken and closed lines, ellipses, squares, and circles after 1-, 3-, and 8-second memory intervals. He found (*a*) an increasing negative time error for the broken ellipse; (*b*) a positive time error for the broken square at 1 and 3 seconds decreasing to zero at 8 seconds; (*c*) no time errors for the other figures. Lesser argued for some type of interaction between the standards and the surround. We see in the following section on context effects that Lesser was probably correct. An experiment similar to Lesser's was performed by Crumbaugh (1954), who found a positive time error for a gap in a circle (a widening of the gap) to 5 seconds, followed by an increasing negative function. A similar result was obtained for the angle of tilt in a line. Other results showed that ellipses grew more circular (interpreted as a negative time error) and unequal circles grew first more equal (negative time error) then unequal with longer memory intervals.

Tresselt (1944b) also tested time errors in visual extents. She found large negative time errors for broken lines or circles but no errors when these figures were solid. The author argued that stimulus characteristics are the determining factor for time errors rather than a fading or intensification of a memory trace.

Karlin (1953) presented black solid circles to subjects on white cards. He found an increasing negative *p*-function for the comparison of sizes. The entire curve was raised in a positive direction with increases in presentation time (1, 3, 5 seconds) for the standard.

Andrews and Robinson (1948) investigated time errors in the Müller-Lyer illusion. They successively presented the two figures involved in the illusion and found a negative *p*-function (decrease in the illusion), which became increasingly positive for memory intervals greater than 6 seconds.

Auditory research. Vance (1914) found that male subjects showed mostly negligible time errors, whereas females had positive or negative time errors, which varied systematically according to the frequency of the standard.

Koester (1945), contrary to Wada but in agreement with Postman (1946), found negligible time errors for pitch. His results indicated that constant errors for loudness judgments are somewhat larger than those for pitch but do not systematically relate in size or direction to interval length.

Pratt (1933b) tested time errors in auditory intensities with an absolute identification method in which subjects rated the standard. Time errors were determined by subtracting the average rating for the series from the mean of the rating scale values. Contrary to Katz's theoretical predictions, Pratt found that loud stimuli produced negative time errors, while positive errors were obtained with soft stimuli.

Other research. Woodrow (1933a) found a large interaction between time error direction and subjects for a temporal spacing task. Subjects were evenly divided between systematic positive and negative *p*-functions; the effect of averaging these subjects was an overall *p*-function close to zero.

Context Viewpoint

Researchers holding a context viewpoint had found the simple trace explanations for time errors such as those of Köhler (1923), Koffka (1935), and Katz (1935), greatly at odds and guilty of gross oversimplifications. None of the theories, for example, explained why apparent fading or intensification effects produce time errors no larger than about 20% instead of the enormous errors that should result if a trace completely faded or greatly intensified. Apparently, traces do not fade completely or intensify greatly, or there are additional mechanisms involved.

The context view holds that a pair of stimuli is not presented in a vacuum; rather, the evaluation of the comparison stimulus is affected by previous stimuli, such as the standard, the level of stimuli in the series, and by the "absolute scale of stimulus meaning prescribed by experience [Osgood, 1953, p. 93]."

Research on context effects may be conveniently divided into two categories: series effects and assimilation effects. Series effects involve time errors that are the result of prior stimulation or of the magnitude of the stimuli in the series. Included here is what Osgood has labeled as the central tendency effect:

> changes in the impression of any standard during the S-V (standard-comparison) interval tend toward the average magnitude of stimuli previously encountered, weak standards augmenting and strong standards fading [Osgood, 1953, p. 97].

We may recall that Katz advanced a trace hypothesis that predicts exactly the reverse. Assimilation effects include time errors attributable to the effect of background stimulation or stimuli interpolated between the standard and comparison stimuli.

Series effects—visual research. Ipsen, according to Osgood (1953), reported a central tendency effect for visual extents, small stimuli producing positive time errors while large stimuli produced negative errors.

Series effect—kinesthetic research. Fernberger (1931) found a large central tendency effect in an experiment employing an absolute judgment method with weights. If light weights were preceded by a heavy weight, they were rated as of intermediate value (a migration toward the preceding stimulus). Preceding heavy by light weights also produced judgments toward an intermediate value. Woodrow (1933b) also employed an absolute judgment method and found that large increases in the magnitude of negative time errors with weights could be obtained by increasing the weight of the standard. Woodrow (1933b) followed his absolute judgment experiment by a carefully controlled study employing successive comparisons. He used standards which were either fixed at 110 or 200 grams or were varied randomly within the same range. A negative time error of 14% was found for each fixed set and as the average of the variable set. However, an analysis within the variable series revealed a large central tendency effect: Positive time errors were obtained for the lightest weights (110–120 grams), zero errors for 130–140 grams, and very large (49%) negative time errors occurred at the high end (190–200 grams). Woodrow concluded (*a*) when stimuli are used in a series a central tendency effect occurs; (*b*) a sinking (fading) of the entire series occurs in the memory. When the standard stimuli are fixed, there is no series effect, and only the sinking effect occurs.

Series effects—auditory research. Needham (1935a) found a marked central tendency effect for sound intensity, which increased with lengthening of the memory interval. He found large negative time errors for 50-decibel and 60-decibel standards, while positive errors were obtained with 30-decibel and 40-decibel stimuli. In another study (1935b), he found that by preceding 30-decibel and 40-decibel standards with a 60-decibel stimulus he could exaggerate the positive time error normally obtained. This result is consistent with Fernberger's finding that standards are drawn in the direction of preceding stimuli. A strong central tendency effect using an absolute judgment method with sound intensity was obtained by Pratt (1933b). He repeated his experiment, but substituted a successive comparison technique; the central tendency effect diminished.

Koester and Schoenfeld (1946) reported an extremely large central tendency effect for pitch (positive around 1000 Hertz [Hz]; negative at 1120 Hz). The effect increased with the length of the memory interval (1 to 6 seconds). The result is interesting in light of the total lack of time errors reported a year earlier by Koester for pitch and loudness based on 7,342 observations. Finally, Long (1937) found that an increase in sound intensity was created in the memory by preceding a loud standard by a soft stimulus. This result conflicts with those of Fernberger and Needham, cited previously, who found the standard drawn toward preceding stimuli.

Assimilation effects—visual research. Marchetti (1942) investigated interpolation and background stimulus effects in the time errors occurring during the comparison of visual extents. He interposed a small (70 millimeter) line between a 100-millimeter standard and a near 100-millimeter comparison stimulus. This resulted in the standard apparently moving in memory toward the interpolation stimulus, thus changing the positive time error that occurred without the interpolated stimulus to a negative value. Background effects were revealed by increasing the size of the cards bearing the standard. Larger cards increased the positive time error (an assimilation toward the background), while decreasing the card size decreased the magnitude of the positive time error.

Freeman and Sharp (1941) reported that Lauenstein found positive time errors when lines drawn on white cards were successively compared to a 100-millimeter standard. When the standard and comparison stimuli were changed to light traces in a darkened room, a slight negative time error was obtained—a result apparently attributable to the removal of the background. This interpretation was supported by McClelland (1943), who varied the size of the frame for comparison stimuli and the 100-millimeter standard. He found an apparent assimilation of the standard to the frame.

Tresselt (1944a) theorized that the effect of background would be greater if the object stimulus was similar (homogeneous) to the background. He compared time errors for black outlined squares in black framed squares (homogeneous condition) with black lines in square black frames (heterogeneous condition); there were also two conditions of size of stimulus objects on the same background sizes. The hypothesis was supported: time errors for the black lines were not significant, while the small black squares produced significant negative time errors and large black squares significant positive time errors.

Crannel (1941), in a visual extent experiment, found results that conflicted with most assimilation experiments. Narrowly and widely spaced dots were interpolated between pairs of dots projected on a screen. Negative time errors (dots closer together) were found for narrow dot interpolations; however, the interpolation of the widely spaced dots produced an even larger negative time error. He concluded that after a certain point assimilation ceases and a contrast effect occurs. Tresselt (1944a) found a similar contrast effect with very small squares on a very large background.

Assimilation effects—kinesthetic research. Guilford and Park (1931) found assimilation of the standard to interpolated weights.

Assimilation effects—auditory research. Lauenstein, according to Pratt (1933a), was one of the earliest proponents of the context position, arguing for a total assimilation theory of time error. He replicated Köhler's click experiment and obtained identical results; however, when he interpolated a loud sound between the standard and comparison stimuli, a large positive time error occurred. Lauenstein argued that Köhler's largely negative p-function occurs because the standard assimilates to the zero stimulus between the standard and comparison stimulus. His own results demonstrated that the principle holds if a background stimulus is presented. A background louder than the standard resulted in a positive time error while a softer background gave rise to a negative error.

Pratt (1933a, 1933b, 1936) tested Lauenstein's hypothesis in a series of studies. He interpolated either loud, soft, or no sound between the standard and comparison stimuli. A positive time error was obtained for the loud interpolation and negative errors for the soft and null interpolations; the time error from the soft interpolation was larger than for the null interpolation, a result supported by similar findings already cited. Pratt argued, as had Woodrow, for a combined trace and assimilation theory where interpolated loud and soft stimuli produce an assimilation in addition to a sinking effect. (When the time errors of the 20% louder and softer interpolated stimuli were averaged, a net negative time error existed.) Only a sinking effect occurs with a null interpolation. Needham (1934a, 1935b, 1935c) also argued for a combined sinking trace-assimilation explanation. He found the effect of an interpolated stimulus wore off during longer memory intervals and an increasing negative p-function inevitably occurred.

Bowland (1958) advocated a revised assimilation position. He contended that traces do not sink; rather, they exist in "persistent groups" that are pulled toward interpolated stimuli. He tested his contention with an experiment in which there was either no interpolated stimulus or one with the same value as the standard; he argued that Pratt should predict a greater sinking trace for the former condition. Bowland found no time errors for either case, and maintained therefore that traces do not sink but only assimilate. However, he is in the dangerous position of accepting the null hypothesis to prove his case; moreover, his lengthy dissertation provides an excellent argument for reviewing the literature—he picked a memory time of 3 seconds which is the exact point in Köhler's p-function where no time error occurred. His results must be considered inconclusive at best.

Quantification of context effects—Helson's adaptation level theory. Helson (1947) extended his adaptation level (AL) theory to cover the effects of series, background, and interpolated stimuli. He theorized that successive comparisons take place within a framework where a neutral point or adaptation level exists. Context stimuli alter the adaptation level affecting the comparison according to the following expression: $\log(\mathrm{AL} + K) = (3 \log X_i/n + \log C)/4$, where K is a constant depending on whether the standard or comparison stimuli are presented first, X_i is the value of each stimulus judged, C is the value of the standard, and n is the number of stimuli in the series. Helson analyzed Pratt's and Fernberger's data and found that his equation described their results exactly. This formulation provides the interesting conclusion that negative time errors may arise because subjects geometrically average standard, comparison, and context stimuli, and as the geometric mean is usually beneath the arithmetic mean, the difference would appear as a negative time error. Helson's formulation does not include time as a parameter; therefore, it covers only one point of the p-function. It would probably not be too difficult to introduce the time variable into the equation to complete the analysis.

Adaptation View

The simplest explanation (Osgood, 1953) for time errors was considered in 1899 by Muller and Martin (according to Needham, 1934a) and by Newhall et al. (1957). Sensory adaptation or fatigue from exposure to the standard could result in two possibilities: (*a*) if the sensory organ was desensitized and the comparison stimulus followed rapidly behind the standard, the comparison stimulus would appear too weak, producing a positive time error; (*b*) if the sensory organ had enough time to recover, the standard, which would be remembered near its adapted level, would appear weaker than the comparison stimulus, and a negative time error would result. This explanation would parallel the typical p-function first found by Köhler but would be difficult to apply to nonintensive stimuli variations such as pitch or hue. Little data have been advanced to separate this alternative from trace or context effects explanations.

Other Factors Affecting Time Errors

Practice. Köhler (1923) found the positive portion of the *p*-function to diminish with practice. Needham (1934b) found a diminution in time errors with practice as well (1934a) as a tendency for the *p*-function to become positive. Karlin (1953) confirmed Needham's results, finding a rapid drop in the size of time errors with practice.

Set. In 1899 Muller and Schumann, according to Woodworth and Schlosberg (1954), theorized that subjects have a set to lift comparison weights with the same effort as required by the standard; thus, if the comparisons required more effort they are judged to be heavier than the standard. They felt close attention to the comparison stimulus would result in an apparent increase in lifting effort—a negative time error.

Several researchers have shown that muscle potentials closely follow the shape of the *p*-function found by Köhler (Freeman & Sharp, 1941; Payne & Davis, 1940). These authors suggested that subjects initially prepare to lift the comparison stimulus with more effort than the standard (positive time error); however, with longer memory intervals they steadily relax their expectations (an increasing negative time error). It has also been suggested that other time error results are related to cochlear and retinal potentials; however, little effort has been expended to investigate the peripheral concomitants of time error.

Absolute versus relative comparisons. Several investigators have utilized an absolute judgment technique to study time errors. Standards are typically assigned rating numbers; the difference between the average rating of the standards (which must be presented an equal number of times) and the mean of the rating scale yields a measure of time error.

Philip's (1940) subjects rated the color mass of test chips. He found a large central tendency effect; the sign and degree of time errors is highly correlated with the actual color masses of the standards. Pratt (1933a) had found, as Philip (1940), a large central tendency effect for sound intensity with an absolute judgment technique, soft tones resulting in positive time errors and loud tones in negative errors. When a successive comparisons method was employed, however, the central tendency effect disappeared.

Woodrow (1933b) and Fernberger (1931) found overall negative time errors for lifted weights with an absolute judgment method; moreover, they found a central tendency effect. Wever and Zener (1928) compared absolute and relative judgment methods for lifted weights. They found a negative *p*-function for both methods; the time error was larger for the absolute method (6% compared to 4% for the relative method). They hypothesized that with time the entire series of weights lifted in the past fades, shifting the subjective scale downward in the negative time error direction. Needham (1935b), taking note of Wever and Zener's results, hypothesized a shift from a relative to an absolute judgment comparison for very long memory intervals in relative comparison experiments.

Stimulus fadeout. McClelland (1943) found time errors were caused in part by the rapidity of stimulus onset and offset. The slow shutoff of projectors during a visual extent experiment produced a negative time error. McClelland noted that in most weight-lifting experiments, subjects put the weights down rather than abruptly dropping them. He tested the effect of abruptly dropping the standards and found a slight positive time error (as contrasted with the negative time error usually obtained). It was hypothesized that similar effects could be found in auditory experiments due to speaker fadeout.

The effects of brain injury and inebriation. Weinstein (1955) reported research on the effects of brain injury and inebriation on time errors. Among the research cited by Weinstein, Specht found inebriation drastically increases negative time errors with auditory stimuli. Similar results with visual stimuli were obtained by Lange and Specht. Large positive time errors with disorders of sensory functioning were observed by Stein and Wiezsacker (according to Weinstein), while Weinstein found negative time errors which increased with longer memory times for persons who had lesions of the parietal lobe.

Psychophysical methods. The methods of limits, constant stimuli, absolute judgment, and average error have most frequently been used to determine time errors. The researchers on color memory have favored the average error technique, while the time error studies have usually been based on the constant stimuli method.

Postman (1947) sought to determine the effect of using the average error method as compared to the constant stimuli approach. He replicated Köhler's study several times, employing the constant stimuli technique; he obtained Köhler's results each time. When he substituted an average error method, he was unable to find any negative time errors and only slight positive time errors. Postman noted an important difference between the methods —matching time. When the constant stimuli method is employed, maintaining times normally are from 1 to 4 seconds. On the other hand, matching times for the average error method may, as in Newhall et al.'s (1957) study, be greater than 30 seconds. The lengthy matching time for the average error technique must be suspect, especially when one considers that the comparison stimulus, which is under adjustment throughout the match interval, acts basically as a long duration interpolated stimulus.

SUMMARY OF MEMORY TIME RESULTS

Direction and Magnitude of Time Errors

Studies involving the effects of memory time may be separated into two categories based on the type of stimulus dimension varied: (*a*) Quantitative, where an intensity dimension is varied that alters such phenomenal attributes as loudness, brightness, saturation, and weight; and (*b*) qualitative, where variation of the stimulus dimension changes such phenomenal attributes as hue, pitch, visual extent, and geometric configuration. This classification is similar to S. S. Stevens' (1957) distinction between prothetic and metathetic continua. An attempt was made to sort the results of studies involving the possibility of time errors into these categories and according to the predominant trend of the results: no time errors, negative time errors, positive time errors, approximately equal positive and negative time errors, and evidence that the time error depended on context variables. The tabular results are presented in Table 1. (A tally was made for each result, not just for each study; classifications are sometimes difficult to make according to these crude categories.)

TABLE 1

TABULAR BREAKDOWN OF MEMORY TIME RESULTS

Predominant p-function trend	Time error research		Color memory research	
	Stimulus dimension varied			
	Quantitative	Qualitative	Quantitative	Qualitative
No error	1	0	3	3
Negative	12	4	2	0
Positive	2	4	10	1
Equal positive and negative	1	4	0	1
p-function depends on context	17	9	—	—
Total	33	21	15	5

We may conclude from Table 1 that the time error literature indicates the prevalence of a negative p-function where quantitative stimulus dimensions were varied; however, a strong interaction with context is obvious. This conclusion is in conflict with the general trend of the color memory research. No clear trend seems to exist for time errors with qualitative variations in stimuli; however, stimulus context appears to be equally important for qualitative as for quantitative variation of stimuli.

The Effect of Context

1. Central tendency. Nine studies tested for a central tendency effect; of these, eight observed it. Models hypothesizing the migration of memory traces toward a norm appear to be useful.

2. Previous stimuli. Two of the three studies reporting the effect found an apparent assimilation of memory traces to the mean of stimuli presented prior to the standard.

3. Background. Five studies were conducted to determine the effect of background. All revealed an apparent assimilation of the memory trace to background stimuli.

4. Interpolated stimuli. Of the nine studies investigating interpolation, eight indicated an apparent assimilation of the memory trace to

the interpolated stimulus; the dissenting study reported a contrast effect. Assimilation apparently occurs until the differences become so large that they trigger a contrast effect.

Other Considerations

1. Practice appears to reduce time errors. There was some indication that highly practiced subjects may make significantly more positive errors.

2. There may be a peripheral component to time errors. Several studies reported muscle potentials correlated with time error *p*-functions.

3. Absolute comparisons may produce more negative errors than relative comparisons; however, data are sparse on this point.

4. McClelland presented evidence that indicated the rapidity of the standard stimulus offset may greatly affect time errors.

5. Brain injury or inebriation resulted, in three of four cases, in larger negative time errors. A tentative explanation would be a change in the memory store capacity.

6. Postman found large differences between constant stimuli and average error psychophysical techniques. Large negative errors were obtained with the former, while a slightly positive *p*-function was obtained with the latter method.

7. Color memory researchers obtained different results than time error researchers. With quantitatively varied stimuli, the former found large positive time errors while the latter obtained increasing negative functions. (The trend for qualitatively varied stimuli is not clear for either group.) Several factors arising from the different experimental approaches of the two groups, or from poor techniques, appear potentially responsible for the disparate results.

CRITIQUE FOR RESEARCH ON MEMORY TIME EFFECTS ON SUCCESSIVE JUDGMENTS

Consideration of the research reviewed in this study has led us to make the following criticisms and recommendations for studies involving the effects of memory time in successive judgments:

1. *Control of context stimuli*. Many experimenters have neglected to control for context stimuli. This may be particularly important in color judgments. Little concern has been shown for background brightnesses and homogeneity, stimulus intensity, or stimuli preceding or following the standard.

2. *Adaptation and afterimages*. Few controls have been introduced for the effects of sensory adaptation or afterimages (Peak, 1939). For short memory intervals the effects of adaptation and afterimages would be identical in color memory—both would give rise to positive time errors. As noted earlier, adaptation might lead to negative time errors for longer memory times. Strong interpolated or preceding series stimuli probably produce some adaptation and moderate afterimages, resulting in an apparent assimilation toward the intervening stimuli.

The color memory results were probably confounded with adaptation and afterimage effects. Presentation times (for the standard) tended to be lengthy (up to 15 seconds), probably producing moderate adaptation and vivid afterimages. Moreover, use of the average error technique, by maintaining color stimuli during the memory interval, undoubtedly led to large adaptation and afterimage effects; the large positive *p*-functions obtained by the color memory researchers bolster these contentions.

Sperling (1960) noted that by irradiation afterimages might alter nonintensive characteristics such as object size.

3. *Practice*. Scant attention has been paid to practice effects by either color memory or time error researchers. The small time errors reported by many investigators might be the result of practice.

4. *Analytical factors*. (*a*) Psychophysical method. Cited earlier were data obtained by Postman that indicated large differences between the methods of average error and constant stimuli techniques. Since the majority of color memory research involved use of the average error technique while time error researchers employed the constant stimuli method, the apparently discrepant results must in part be attributed to this difference. Siegal (1962) found the constant stimuli method to be the most sensitive (smallest discrimination threshold values) of the psychophysical techniques. Moreover, we have noted the average error technique is suscepti-

ble to contamination from stimulus adaptation and afterimages. Large assimilation effects (toward the origin of the series) have been reported from the few studies using the method of limits. The constant stimuli method, therefore, appears best for time error research.

(*b*) Calculation of time errors. Several methods exist for calculating time errors for the constant stimuli techniques: The simple E% method, freehand and sliding average ogive fits, normalized straight line fits, as well as weighted and unweighted normalized least squares fits, can be utilized (Guilford, 1954). The time error and color memory studies have used many of these techniques without concern; it is inconceivable that no differences exist among them; moreover, the results may not be comparable to the time errors obtained through average error or limits methods.

(*c*) Individual and sex differences—the consequences of averaging. Little attention has been focused upon the question of individual or sex differences. Woodrow (1933b) reported an even split for subjects: half had positive *p*-functions and half, negative *p*-functions. No evidence has been presented to show the sexes are equivalent. With these questions unresolved, averaging has obvious dangers. Techniques devised for the practical problem of reducing time errors may not hold if based upon averages; individual selection and training may be more fruitful for this practical purpose.

(*d*) Statistical evaluation. Little of the research on color memory or time error has been subjected to the scrutiny of statistical evaluation; usually, only mean time errors have been reported. There is actually little evidence to discredit the argument that each body of evidence is actually the product of guessing from subjects biased toward a particular result. Only the frequency of similar conclusions leads us to a contrary belief.

(*e*) Sample size. Most of the studies cited have utilized very few subjects (which is probably the reason few statistical analyses have been attempted). Hamui and Landis (1955) used two subjects; Newhall et al. (1957), three; Payne and Davis (1940), only one. Bowland (1958) used 26 subjects, but gave each only three presentations of a specific stimulus. Sample sizes as small as these are hardly reliable enough to make the broad generalizations typically made by the investigators.

(*f*) Weber's law. Except for Helson's treatment, there has been a total disregard of the relativity of stimulus change quantified by Weber's law. Changes in low-intensity stimuli have been treated equally to changes in stimuli of high intensity. We must relate the magnitude of a stimulus change to the absolute stimulus level. "To dismiss this law because it does not hold over the entire range would be to obscure a fundamental fact about sensory mechanisms [Osgood, 1953, p. 76]." Averaging large and small stimuli, divided by their absolute magnitude, may no longer result in an overall negative time error as has been previously shown. Many of the central tendency effects are undoubtedly contaminated with this error.

(*g*) Guessing and response bias. Little attention has been paid to the problem of contamination from guessing and subject bias. If a subject cannot sense differences among the comparison stimulus alternatives and is forced to guess, his *p*-function will appear as a horizontal line centering close to a zero time error. Response biases will move the *p*-function markedly above or below zero. It is difficult to estimate what proportion of the color memory and time error results are confounded by guessing and bias, but it is interesting to note the tendency for the *p*-function to go to zero at long intervals where subjects might be expected to rely most heavily on guessing (Karlin, Lehmann, Wolfe, Whipple, Wada; in Needham, 1934b).

5. *Fatigue.* Bowland (1958) noted that experimenters have typically given little consideration to the length of their experiments; time error studies are extremely fatiguing. Fatigued subjects cannot be expected to provide adequate data.

6. *Fixed versus variable standards.* In the majority of the studies, a single fixed standard has been repeatedly presented. In real life, observers are often called upon to make comparisons with standards that are different from one judgment to another. In light of practice and series effects, it is probable that large differences exist between judgments

made with fixed standards and those made with variable standards.

7. *Stimulus fadeout.* As McClelland (1943) found, large time error differences can result by varying the rapidity of removing the standard. It is reasonable to speculate that both onset and offset of both standard and comparison stimuli may have considerable effect on time errors. These variables have been largely uncontrolled.

MODEL STUDY OF SATURATION MEMORY

In a series of carefully controlled studies of saturation memory for a single 538-millimicron stimulus, Springer and Tate[3] attempted to take these problems into account. A constant stimuli method was employed with memory intervals of 1, 4, 14, or 30 seconds. There were several parts to the study, in which the degree of saturation, standard variability, and exposure time of the standard stimulus were varied, and the effects of practice, individual, and sex differences were assessed. In addition, the results of different psychophysical curve-fitting techniques were compared. Precise controls were exercised over the color purity and brightness level of the stimuli. When the data for all subjects were analyzed, no significant time errors (via sign tests) or differences among conditions (via analysis of variance) were found. However, a consideration of the individual *p*-functions indicated that part of the difficulty in such studies may be due to subjects who have very poor saturation perception and hence do a lot of guessing when asked to make such judgments. Such indiscriminability would show up in large difference limens. Typical limen values ranged from 9% to 12%, and it was decided to reject all subjects having a difference limen of 16% or greater; this resulted in 16 of 60 data cells being discarded. This procedure is certainly reasonable if we are interested in saturation memory among those able to make good saturation discriminations. An analysis of variance of these "corrected" data did show significant differences among the different memory intervals, but no other significant main effects or interactions. The data showed a *p*-function with a slightly positive time error at the 1-second memory interval, which changed to zero at 5 seconds, and became more and more negative at the 15- and 30-second intervals. This finding supports the results of previous time error studies on quantitatively varied stimuli, but conflicts with color memory experimental results. Both the 15- and 30-second memory groups were significantly different from zero (i.e., significant time errors). Not only were the analyses significant where they had not been before, but differences in *p*-functions resulting from different curve-fitting techniques were less for the corrected than the uncorrected data.

Other findings of the studies were that practice with feedback reduced the time error to insignificant amounts for all conditions. Of the individual "reliable" subjects, 13 gave more positive time errors with practice. It was also found that males had significantly more positive time errors than females in 11 of 16 experimental conditions. Another sex difference was that women tended to be more accurate at the shorter memory times while males were more accurate at the longest intervals used. The fact that in the unreliable data more subjects tended toward positive time errors may be a clue to tendencies noted in studies by Karlin (1953) and others for time errors to shift in a positive direction at longer memory intervals. At longer time intervals, one would expect guessing and response biases to be more prevalent and hence beget unreliable data.

CONCLUSION

The Springer and Tate study (see Footnote 3) clearly demonstrates the necessity of taking into account factors mentioned in the preceding critique, at least in the study of the effects of memory time on successive judgments of color. However, the results and the critique have broader implications for psychophysical methodology than just for study of errors in successive judgments. It is apparent that the exercise of careful controls of a number of variables is a very critical matter in any study involving a psychophysical procedure; the fact that such procedures are part of the oldest branch of experimental psychology methodology does not mean that we have yet learned, much less practiced, all the refinements desirable in these techniques. Our analyses of psychophysical data have also left much to be desired. Certainly no study should ignore, as they usually do, considerations of the effects of individuals and sex differences and the inclusion of poor discriminators in the data; such data can give confusing and misleading results.

---

[3] R. M. Springer and J. D. Tate. Article in preparation.

REFERENCES

ALLPORT, G. W. Change and decay in the visual memory image. *British Journal of Psychology*, 1931, **21**, 133–148.

ANDREWS, T. G., & ROBINSON, I. Time error and the Müller-Lyer illusion. *American Journal of Psychology*, 1948, **61**, 224–235.

BARTLESON, C. J. Colors of familiar objects. *Journal of the Optical Society of America*, 1960, **50**, 73–77.

BENTLEY, I. M. The memory image and its qualitative fidelity. *American Journal of Psychology*, 1899, **11**, 1–48.

BORING, E. G. *Sensation and perception in the history of experimental psychology*. New York: Appleton-Century-Crofts, 1942.

BOWLAND, J. A. *The effect of varying intensity on the auditory time error.* (Doctoral dissertation, University of Florida) Ann Arbor, Mich.: University Microfilms, 1958, No. 58-3481.

BURNHAM, R. W., & CLARK, J. R. A test of hue memory. *Journal of Applied Psychology*, 1955, **39**, 164–172.

CHRISTAL, R. E. Factor analytic study of visual memory. *Psychological Monographs*, 1958, **72**(13, Whole No. 466).

COLLINS, M. Some observations on immediate colour memory. *British Journal of Psychology*, 1931, **22**, 344–352.

CRANNELL, C. W. The effect of interpolated stimuli on the time error in judgments of vertical distance. *Psychological Bulletin*, 1941, **38**, 609.

CRUMBAUGH, J. C. Temporal changes in the memory of visual perceived form. *American Journal of Psychology*, 1954, **67**, 647–658.

DAVIS, R. C. The stimulus trace in effectors and its relation to judgment responses. *Journal of Experimental Psychology*, 1952, **44**, 377–390.

FERNBERGER, S. W. On absolute and reflective judgments in lifted weight experiments. *American Journal of Psychology*, 1931, **43**, 560–578.

FREEMAN, G. L., & SHARP, L. H. Muscular action potentials and the time-error function in lifted weight judgments. *Journal of Experimental Psychology*, 1941, **29**, 23–36.

GUILFORD, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954.

GUILFORD, J. P., & PARK, D. G. The effect of interpolated weights upon comparative judgments. *American Journal of Psychology*, 1931, **43**, 589–599.

HAMUI, V., & LANDIS, C. Memory for color. *Journal of Psychology*, 1955, **39**, 183–194.

HANAWALT, N. G., & POST, B. E. Memory trace for color. *Journal of Experimental Psychology*, 1942, **30**, 216–227.

HELSON, H. Adaptation-level as frame of reference for prediction of psychophysical data. *American Journal of Psychology*, 1947, **60**, 1–29.

IRWIN, F. W., & SEIDENFELD, M. A. The application of the method of comparison to the problem of memory changes. *Journal of Experimental Psychology*, 1937, **21**, 363–381.

KARLIN, L. The time error in the comparison of visual size. *American Journal of Psychology*, 1953, **66**, 564–573.

KATZ, D. *The world of color*. London: Kegan-Paul, 1935.

KOESTER, T. The time error and sensitivity in pitch and loudness discrimination as a function of time interval and stimulus level. *Archives of Psychology, New York*, 1945, No. 297.

KOESTER, T., & SCHOENFELD, W. N. The effect of context upon judgments of pitch differences. *Journal of Experimental Psychology*, 1946, **36**, 417–430.

KOFFKA, K. *Principles of gestalt psychology*. London: Kegan-Paul, 1935.

KÖHLER, W. Zur Theorie des sukzessivvergleichs und der Zeitfehler. *Psychologische Forschung*, 1923, **4**, 115–175.

KREEZER, G. The neurological level of the factors underlying time error. *American Journal of Psychology*, 1938, **51**, 18–43.

LESSER, J. The effects of shape and closure on the time error. *Dissertation Abstracts*, 1959, **20**, 1081.

LONG, L. A study of the effects of preceding stimuli upon judgments of auditory intensity. *Archives of Psychology, New York*, 1937, No. 209.

MARCHETTI, P. V. Time errors in judgments of visual extents. *Journal of Experimental Psychology*, 1942, **30**, 257–261.

MCCLELLAND, D. C. Factors influencing the time error in judgments of visual extent. *Journal of Experimental Psychology*, 1943, **33**, 81–95.

NEEDHAM, J. G. The time error in comparison judgments. *Psychological Bulletin*, 1934, **31**, 229–241. (a)

NEEDHAM, J. G. The time error as a function of continued experimentation. *American Journal of Psychology*, 1934, **46**, 558–567. (b)

NEEDHAM, J. G. The effect of time interval upon time error at different intensive levels. *Journal of Experimental Psychology*, 1935, **18**, 530–543. (a)

NEEDHAM, J. G. Contrast effects in judgments of auditory intensities. *Journal of Experimental Psychology*, 1935, **18**, 214–226. (b)

NEEDHAM, J. G. Interpolated effects with different time intervals. *Journal of Experimental Psychology*, 1935, **18**, 767–773. (c)

NEWHALL, S. M. Effects of attention on the intensity of cutaneous pressure and on visual brightness. *Archives of Psychology, New York*, 1932, No. 61.

Newhall, S. M., Burnham, R. W., & Clark, J. R. Comparison of successive with simultaneous color matching. *Journal of the Optical Society of America*, 1957, **47**, 43–56.

Osgood, C. E. *Method and theory in experimental psychology*. New York: Oxford Press, 1953.

Payne, B., & Davis R. C. The role of muscular tension in the comparison of lifted weights. *Journal of Experimental Psychology*, 1940, **27**, 227–242.

Peak, H. Time order error in successive judgments and in reflexes. I. Inhibition of the judgment and the reflex. *Journal of Experimental Psychology*, 1939, **25**, 535–565.

Philip, B. B. Time errors in discrimination of color mass by the ranking method. *Journal of Experimental Psychology*, 1940, **27**, 285–302.

Postman, L. The time error in auditory perception. *American Journal of Psychology*, 1946, **59**, 193–219.

Postman, L. Time error as a function of the method of experimentation. *American Journal of Psychology*, 1947, **60**, 101–108.

Pratt, C. C. Time errors in the method of single stimuli. *Journal of Experimental Psychology*, 1933, **16**, 798–814. (a)

Pratt, C. C. The time error in psychological judgments. *American Journal of Psychology*, 1933, **45**, 292–297. (b)

Pratt, C. C. The law of disuse. *Psychological Review*, 1936, **43**, 83–93.

Siegel, M. H. Discrimination of color. I. Comparison of three psychological methods. *Journal of the Optical Society of America*, 1962, **52**, 1067–1070.

Sperling, G. The information available in brief visual presentations. *Psychological Monographs*, 1960, **74**, (11, Whole No. 498).

Stevens, S. S. On the psychophysical law. *Psychological Review*, 1957, **64**, 153–181.

Tresselt, M. E. Time errors in successive comparisons of simple visual objects. *American Journal of Psychology*, 1944, **57**, 555–558. (a)

Tresselt, M. E. The time errors in visual extents and areas. *Journal of Psychology*, 1944, **17**, 21–30. (b)

Vance, T. F. Variation in pitch discrimination within the tonal range. *Psychological Monographs*, 1914, **16** (3, Whole No. 69), 115–149.

Warren, H. C., & Shaw, W. T. Further experiments on memory for square size. *Psychological Review*, 1895, **2**, 239–244.

Weinstein, S. Time error in weight judgment after brain injury. *Journal of Comparative and Physiological Psychology*, 1955, **48**, 203–207.

Wever, E. G., & Zener, K. E. The method of absolute judgment in psychophysics. *Psychological Review*, 1928, **35**, 466–492.

Woodrow, H. Individual differences in the reproduction of temporal intervals. *American Journal of Psychology*, 1933, **45**, 271–281. (a)

Woodrow, H. Weight-discrimination with a varying standard. *American Journal of Psychology*, 1933, **45**, 391–416. (b)

Woodworth, R. S., & Schlosberg, H. *Experimental psychology*. (Rev. ed.) New York: Holt, Rinehart & Winston, 1954.

(Received March 24, 1970)

Psychological Bulletin
1971, Vol. 76, No. 6, 409–410

COMMENT ON ALTRUISM:

THE COMPARATIVE EVIDENCE

D. O. HEBB[1]

McGill University

Krebs' dismissal of the evidence of altruism in infrahuman mammals is criticized. It is proposed that this is in fact the only decisive evidence, in view of the difficulty of interpreting the human data.

Krebs (1970) has obscured the picture of altruism. He dismisses the comparative evidence as mostly anecdotal, which it is not, and though he agrees that one animal may give help to another he feels that no conclusions are warranted because "the reasons behind [such behavior] are almost as ambiguous as those behind the behavior of humans [p. 262]." Almost, but not quite: The fact is that the only decisive evidence of the existence of altruism, as such, comes from the large-brained infrahuman mammals.

The primary problem is whether there is purposive behavior in which one animal helps another at some cost to himself (in material wealth, such as food, or in the expenditure of energy), when the behavior has not been induced by reward or by threat of punishment. Superficially, such behavior is common in man, but no conclusive argument based on human data has been possible because of the complexity of human experience and all the situations in the past in which the child has been rewarded for giving or scolded for not giving. Here the animal data are decisive in showing that altruism can occur. Nothing other animals do proves anything about man, but such evidence does refute the principal criticism of the human data: that doctrinaire argument, that altruism in the true sense cannot exist, that all motivation must be selfish (based ultimately on primary biological needs).

The idea that the evidence of Hebb and Thompson (1954) was anecdotal is absurd. We referred to the behavior of a dog that prevented children from swimming, by persistently "rescuing" them despite never having been taught anything of the kind and despite repeated punishment, as an example of the well-known guarding and protective behavior often seen in dogs, and Krebs perhaps considers this to be anecdotal. Our main evidence, however, was experimental. We quoted Nissen on the experiments of Nissen and Crawford (1936) and reported some results from an experiment by Hebb (1949). The latter provided a clear example of concern by one adult chimpanzee for a friend's welfare, shown by attempting to pull her away from a place of risk.

Krebs further implies that unsolicited help or food giving is rare in the chimpanzee, and that if the help has been solicited, it is not altruism. The first is not so, and the second is hard to understand. Nissen and Crawford reported a number of examples of unsolicited food giving between the young animals of their experiment, and it is also recorded in the diaries of adult animals of the Yerkes Laboratories of Primate Biology.[2] On the second point, is it not generosity to give food when requested? It may be *more* generous to anticipate another's need (as Nissen and Crawford's half-grown animals in fact did), but giving without return is altruism, whether induced by begging or not. Krebs emphasizes also that the animals gave the less attractive food; but this, instead of casting doubt on the existence of altruism, just makes the case more credible.

Finally, it appears that some of the food giving was imitative: "a result of modeling [Krebs, 1970, p. 262]." This suggests one source of altruism, but altruism does not become less altruistic when we understand how it comes about. In my own opinion, empathy is involved (for evidence of chimpanzee empathy, see Köhler, 1927, Plate IV), as it is in imitation, so it is quite reasonable to find an interaction of altruism and imitation. It is clear, however, from Nissen and Crawford's data and from my own experimental observations (referred to above) that altruistic behavior can occur in the absence of any model to imitate.

---

[1] Requests for reprints should be sent to the author, Department of Psychology, McGill University, Montreal 101 Quebec, Canada.

[2] Now The Yerkes Regional Primate Research Center, Emory University, Atlanta, Georgia.

As I have said, finding that something is true of other animals does not prove it true of man; but if there are suggestions of altruism in the dog, and clear evidence of it in larger-brained animals such as chimpanzee and porpoise (Kellogg, 1961; Siebenaler & Caldwell, 1956), it suggests at least that we might look for it in another large-brained animal—man. In fact, it is well known in man; but it has traditionally been explained away as a product of childhood training on the basis of the classic assumption that the only unlearned motivations must be selfish. The comparative evidence is crucial in one respect: It shows that altruism may arise in the chimpanzee, in some modest degree, where there has been no training in generosity. On any reasonable view, this requires a reinterpretation of the traditional hedonistic, law-of-effect view of human nature and human motivation.

REFERENCES

Hebb, D. O. Temperament in chimpanzees: I. Method of analysis. *Journal of Comparative and Physiological Psychology*, 1949, **42**, 192–206.

Hebb, D. O., & Thompson, W. R. The social significance of animal studies. In G. Lindzey (Ed.), *Handbook of social psychology*. Vol. 1. Cambridge: Addison-Wesley, 1954.

Kellogg, W. N. *Porpoises and sonar*. Chicago: University of Chicago Press, 1961.

Köhler, W. *The mentality of apes*. New York: Harcourt, Brace, 1927.

Krebs, D. L. Altruism—an examination of the concept and a review of the literature. *Psychological Bulletin*, 1970, **73**, 258–302.

Nissen, H. W., & Crawford, M. P. A preliminary study of food-sharing behavior in young chimpanzees. *Journal of Comparative Psychology*, 1936, **22**, 383–419.

Siebenaler, J. B., & Caldwell, D. K. Cooperation among adult dolphins. *Journal of Mammalogy*, 1956, **37**, 126–128.

(Received May 5, 1970).

Psychological Bulletin
1971, Vol. 76, No. 6, 411-414

INFRAHUMAN ALTRUISM

DENNIS KREBS[1]

Harvard University

Hebb's critique of Krebs' (1970) review of infrahuman research on altruism is examined. It is suggested that although there is a good deal of descriptive evidence for helping behavior in infrahumans, experimental support for the assumption that the helping responses were based on altruistic motivation is lacking. It is concluded that evidence from human research on empathy has gone furthest to explain altruism.

D. O. Hebb's (1971) critique of Krebs' (1970) review of the comparative evidence for altruism raised a number of issues that warrant examination. Hebb addressed the problem that traditionally has been most closely associated with the study of altruism—the problem of its existence. Hebb argued that the conclusion reached by Krebs (1970)—that comparative research has not established the existence of altruism in infrahumans—was misleading. (It should be emphasized that Krebs, 1970, did not *dismiss* the evidence.) Hebb went on to argue that the only decisive evidence for the existence of altruism comes from research on large-brained infrahuman mammals. The evidence adduced by Hebb, however, was mostly descriptive. Without exception, it failed to explore the motivational bases of the seemingly altruistic behavior it described. In fact, recent research on humans—research concerned with the effect of empathic reactions on altruistic behavior—has gone furthest to explain altruism.[2]

Before discussing altruism, its criteria should be clearly specified. The criteria employed by Hebb require that the motivation for the behavior in question be "intrinsic" (Hebb & Thompson, 1954) and that the behavior not be "induced by reward or by threat of punishment [Hebb, 1971, p. 409]." Intrinsic seems to imply innate. However, Hebb and Thompson (1954, p. 542) entertained the notion that altruism might be a "learned drive." But inasmuch as learned drives are a function of reinforcement (Miller, 1951), the first criterion is inconsistent with the second.

The separation of learned and innate components of behavior has presented a classic problem for psychologists. Although the study of infrahuman animals has supplied one way of researching the problem, none of the studies reported by Hebb (1971) addressed the issue. Hebb and Thompson (1954) seemed to suggest that the phyletic increase in aiding behavior is due to a corresponding increase in an innate predisposition toward altruism. The increase in helping, however, could be due to any of a number of the differences among the animals in question, especially differences in learning experiences.

Hebb's second criterion forms the basis of his argument, but, as pointed out earlier (Krebs, 1970, p. 260), the criterion is overly restrictive. It is logically impossible to establish the absence of reinforcement, because it entails proof of the null hypothesis. Even though infrahumans are not exposed to formal socialization practices, the conclusion that their helping behavior is not a function of reinforcement is not warranted. It seems quite likely that both humans and infrahumans receive rewards for helping and punishments for not helping, quite independent of formal socialization.

Several points need to be made about the evidence adduced by Hebb in support of his conclusion that infrahumans behave altruistically. To begin with, Hebb and Thompson (1954) recognized the deficiencies of anecdotal evidence: "There is a great deal of anecdotal material to suggest the existence of altruism and cooperation in primate societies

---

[1] Requests for reprints should be sent to the author, Department of Social Relations, William James Hall, Harvard University, Cambridge, Massachusetts 02138.

[2] Krebs, D. L. Empathically-experienced affect and altruism. Unpublished doctoral dissertation, Harvard University, 1970.

. . . , but the evidence in general has been so poor that Zuckerman (1932) had to conclude that no altruism exists . . . [p. 541]."

Hebb and Thompson (1954) then went on to present what they described as "experimental" support for the suggestion that chimpanzees and porpoises behave altruistically. None of the studies, however, supplied experimental support for the existence of *altruism*, if experimental means the manipulation of independent variables and the measurement of their effect on dependent variables and altruism means nonreinforced aiding responses. Rather, the evidence was of a descriptive nature; it was based on controlled observation of laboratory animals. The only study that focused on a dependent variable relevant to the understanding of altruism (Nissen & Crawford, 1936) was careful not to extend its conclusions beyond what it examined—food sharing in chimpanzees.

Turning to the studies in question, although the protective behavior of the chimpanzee reported by Hebb (1949) was observed in an experimental setting, the experiment was actually concerned with the reaction of chimpanzees to threatening and nonthreatening stimuli. It occurred in a chimpanzee during an aggressive exchange between her partner and a provocative experimenter. No attempt was made to specify its motivational base. Protective behavior such as that observed by Hebb (1949) seems most appropriately analyzed as *cooperative* because it helps guarantee the welfare of both protector and protected. Cooperation is different from altruism because it refers to behavior in which two or more parties help one another reach a goal beneficial to both.

Behavior similar to the protective responses of Hebb's chimpanzee is common in humans. Campbell (1965) reviewed research on ingroup solidarity in the face of outgroup aggression. Krebs (1970) treated Campbell's (1965) review as "cogent" support for the notion that altruism is not inconsistent with the theory of evolution.

Studies mentioned by Hebb and Thompson (1954) presented descriptions of the behavior of captive porpoises. Although some of the behavior seemed altruistic, no tests were made of the motivational antecedents of the acts. A further study referred to by Hebb, 1971 (Siebenaler & Caldwell, 1956) was experimental, but, by its authors' description, it was concerned with cooperation among porpoises.

A final study (Nissen & Crawford, 1936) supplied unequivocal evidence that chimpanzees share food with one another. It did not, however, show that food sharing was based on altruistic motivations. The solicited food sharing might have been directed at the termination of a noxious begging response, in much the same way as the bar-press responses of the Lavery and Foley (1963) rats (cf. Krebs, 1970, p. 261). To quote Nissen and Crawford (1936):

> Responsiveness to begging may be (1) a response, probably related to reactions given in answer to calls for help and appeals for sympathy, which gets rid of the presumably unpleasant stimulus of begging by the partner, or (2) an exchange with delayed reward . . . [p. 416].

*Un*solicited food sharing supplies more convincing support for altruism because it does not involve a noxious stimulus, but its underlying motives are also unclear. As concluded by Nissen and Crawford (1936): "the unsolicited passing may be explained as (1) play behavior, (2) generalization of the response to begging . . . (3) exchange with delayed or previously received reward [p. 416]." (And unsolicited food sharing *was* relatively rare. Of all the food available, only about 3% was thrown, and about 9% was passed, and almost all of it was passed on a reciprocal basis—chimpanzees, in effect, traded food on a "'credit' or deferred basis"—see Nissen & Crawford, 1936, p. 415.)

To summarize, it is argued that infrahuman research has failed to supply decisive evidence for the existence of altruism. Although incidences of seemingly altruistic behavior in chimpanzees and porpoises (and dogs and rats) have been described, it must be concluded that "the reasons behind these acts are almost as ambiguous as those behind the behavior of humans [Krebs, 1970, p. 262]." The fact that infrahumans are not formally socialized does not mean that they receive "no training in generosity." There are many sources of reward and punishment besides socializing agents.

Perhaps too much stress has been put on the deficiencies of past studies. What type of evidence is needed to establish the existence of altruistic behavior? Hebb (1971) suggested two sources of altruism—imitation and empathy. The effect of modeling on altruistic responses has been discussed elsewhere (Krebs, 1970, pp. 267–276). The effect of empathy warrants further attention.

Research on empathy supplies a way of experimentally supporting the existence of altruism and helping to explain it. Several studies on rhesus monkeys have reported empathic-like responses, but a recent study on humans has supplied the strongest support for the notion that empathic reactions mediate altruistic responses. Mirsky, Miller, and Murphy (1958) found that rhesus monkeys who observed other monkeys emit fear responses performed a previously extinguished avoidance response. Miller, Banks, and Ogawa (1963) found that signs of anticipatory fear in one of a pair of monkeys served as a cue to his partner to press a lever to defer a shock to both. And, in a similar situation, Miller, Caul, and Mirsky (1967) found that normally reared (but not socially isolated) monkeys were able to learn to avoid shock by watching the facial expression of a normal (but not socially isolated) partner. Of considerable interest was the finding that the avoidance responses of the normal monkeys were accompanied by a psychophysiological sign of emotional arousal—cardiac acceleration. It is possible that the monkeys reacted to the information that they might receive a shock rather than to the distress of their partner. Moreover, the behavior of the monkeys was more cooperative than altruistic. However, a further study found quite clear evidence that distress cues are noxious stimuli to rhesus monkeys. Masserman, Wechkin, and Terris (1964) found that monkeys consistently refrained from pulling a chain that brought them food in order to avoid giving a shock to their partners. Even though the monkeys in the Masserman et al. (1964) study could get food on alternate trials by pulling another chain, the evidence suggests that monkeys react empathically to signs of distress in their fellows and that their empathic reactions mediate altruistic responses. The evidence at the human level, however, is much stronger.

It was suggested earlier that the criterion of altruism need not exclude reinforcement. In particular, it is thought that empathically or vicariously experienced reactions supply reinforcement for behavior that can be considered altruistic. The notion is elaborated elsewhere (see Footnote 2). In short, it suggests that if observers experience the joys and sorrows of others vicariously, then in some cases it may be more rewarding for them to give goods to another (and experience his pleasure vicariously) than to keep the goods (and experience a lesser pleasure directly). In this conception the problem of proving the absence of reinforcement is avoided, and the power of hedonism is brought to bear on an essentially nonhedonistic response. The conception is similar to that of Aronfreed (1968, pp. 138–149).

Three experiments support the notion that empathic reactions can mediate altruistic responses. Aronfreed and Paskel (reviewed by Aronfreed, 1968) and Aderman and Berkowitz (1969) created situations designed to induce empathy and found that subjects exposed to these situations displayed a disposition to sacrifice goods and perform helping responses. The empathic reactions of the altruistic subjects, however, were never checked—the investigators only assumed that their manipulations produced empathic responses. A recent study (see Footnote 2) attempted to induce and measure positive and negative empathic responses and to measure their effect on altruistic behavior. Psychophysiological reactions of observers who watched a performer appear to receive rewards and punishments were measured; then the observers were given a chance to help the performer. It was found that subjects who were led to believe that they were similar to the performer (past studies, e.g., Stotland, 1969, have shown that perception of similarity increases the disposition to empathize) evidenced the greatest psychophysiological reactions to his plight (as measured by changes in heart rate, blood pulse volume, and skin resistance). They also reported experiencing most empathy, and they behaved most altruistically. The results were interpreted as supplying experimental support for the notion that humans vicariously experience the reactions of similar others and that these vicarious reactions supply reinforcement for altruistic behavior. It should be clear that even though the altruism in question is a function of reinforcement, the results do not imply a hedonistic self-oriented view of human nature or human motivation. If anything, they suggest a revision of the materialistic conception of human nature based on conceptions of reinforcement popularized by infrahuman research.

In conclusion, it was argued that descriptive evidence is not a sufficient basis for conclusions about altruism, in infrahumans or humans. None of the studies reviewed by Hebb and Thompson (1954) established the existence of altruism in infrahumans or took steps to explain altruistic behavior. Hebb's basic notion—that research on infrahumans supplies a unique way of gaining information about altruism—seems well taken, and infrahuman mammals certainly appear to behave altruistically. It is hoped that future research will investigate the motivational base of infrahuman helping behavior.

REFERENCES

Aderman, P., & Berkowitz, L. Empathy, outcome, and altruism. *Proceedings of the 77th Annual Convention of the American Psychological Association*, 1969, **4**, 379–380.

Aronfreed, J. *Conduct and conscience*. New York: Academic Press, 1968.

Campbell, D. T. Ethnocentric and other altruistic motives. *Nebraska Symposium on Motivation*, 1965, **13**, 283–311.

Hebb, D. O. Temperament in chimpanzees: 1. Method of analysis. *Journal of Comparative and Physiological Psychology*, 1949, **42**, 192–206.

Hebb, D. O. Comment on altruism: The comparative evidence. *Psychological Bulletin*, 1971, **76**, 409–410.

Hebb, D. O., & Thompson, W. R. The social significance of animal studies. In G. Lindzey (Ed.), *Handbook of social psychology*. Vol. 1. Cambridge: Addison-Wesley, 1954.

Krebs, D. L. Altruism—an examination of the concept and a review of the literature. *Psychological Bulletin*, 1970, **73**, 258–302.

Lavery, J. J., & Foley, P. J. Altruism or arousal in the rat? *Science*, 1963, **140**, 172–173.

Masserman, J. H., Wechkin, S., & Terris, W. "Altruistic" behavior in rhesus monkeys. *American Journal of Psychiatry*, 1964, **121**, 584–585.

Miller, N. E. Learnable drives and rewards. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley, 1951.

Miller, R. E., Banks, J. H., & Ogawa, H. Role of facial expression in "cooperative-avoidance conditioning" in monkeys. *Journal of Abnormal and Social Psychology*, 1963, **67**, 24–30.

Miller, R. E., Caul, W. F., & Mirsky, I. F. Communication of affects between feral and socially isolated monkeys. *Journal of Personality and Social Psychology*, 1967, **7**, 231–239.

Mirsky, I. A., Miller, R. E., & Murphy, J. V. The communication of affect in rhesus monkeys: I. An experimental method. *Journal of the American Psychoanalytic Association*, 1958, **6**, 433–441.

Nissen, H. W., & Crawford, M. P. A preliminary study of food-sharing behavior in young chimpanzees. *Journal of Comparative Psychology*, 1936, **22**, 383–419.

Siebenaler, J. B., & Caldwell, D. K. Cooperation among adult dolphins. *Journal of Mammalogy*, 1956, **37**, 126–128.

Stotland, E. Exploratory investigations of empathy. In L. Berkowitz (Ed.), *Advances in experimental social psychology*. Vol. 3. New York: Academic Press, 1969.

(Received June 11, 1970)

Psychological Bulletin
1971, Vol. 76, No. 6, 415–431

PARTIAL REINFORCEMENT:

A SELECTIVE REVIEW OF THE ALLEYWAY LITERATURE SINCE 1960[1]

DONALD ROBBINS[2]

The Rockefeller University

Studies using rats in an alleyway are reviewed with regard to the partial reinforcement acquisition and extinction effects. The magnitude of these effects is complicated by the fact that certain variables have differential effects depending on whether the rat is given a partial or a continuous reinforcement schedule. Three major theoretical positions, dissonance, frustration, and habituation of competing responses, are shown to encounter difficulties with a considerable body of data. It is suggested that extensions and modifications of Capaldi's sequential theory, which handled most aspects of the extinction effect, may prove to be a viable theoretical alternative.

Since the publication of Lewis' (1960) literature review, updating the Jenkins and Stanley (1950) review, there have been an extremely large number of studies investigating the effects of partial reinforcement. This review is even more selective than that of Lewis in that only alleyway studies with rats, using positive reward, are dealt with. The review is divided into three sections: (*a*) a literature review of the partial reinforcement extinction effect (PREE); (*b*) a review of the data on the acquisition effect (PRAE); and, finally, (*c*) an evaluation of theoretical proposals. Because of the voluminous literature, references are made to review and summary articles whenever possible. Occasionally, references are made to articles published prior to 1960, either because they were published too late to be included in Lewis' review or because of their historical significance. It should be noted that the dependent variable in these studies is almost always a speed measure (e.g., start, alley, goal speed), although trials to criterion is occasionally used. Furthermore, unlike Lewis' (1960), the present review does not discuss the effects of intertrial interval separately, but rather, when pertinent, in relation to other variables.

EXTINCTION EFFECT

The basic reference is Weinstock (1954, 1958), who gave different groups of rats one trial per day, under a 17%, 33%, 50%, 67%, 83%, or 100% reinforcement schedule. He found an inverse relation between resistance to extinction and percentage of reinforced trials during training.

Amount of reinforcement. A number of investigators have reported a larger PREE for groups receiving large, in contrast to small, amounts of reinforcement during training (Gonzalez, Roberts, & Bitterman, 1966; Hulse, 1958; Padilla, 1967; Roberts, 1969; Wagner, 1961), although others have reported no difference in the magnitude of the effect (Amsel, Surridge, & Hug, 1969). These differences are complicated by findings (as reviewed by Sperling, 1965a, 1965b) that resistance to extinction and amount of reinforcement are inversely related for groups receiving continuous reinforcement but, for partially reinforced subjects, are either positively related (Hulse, 1958; Wagner, 1961) or not related (Brooks & Dufort, 1967; Roberts, 1969). The greater magnitude of the PREE for larger amounts of reinforcement apparently results from a combination of a decrease in resistance to extinction for continuously reinforced groups and either an increase or no effect on partially reinforced subjects.

---

[1] This research was supported by United States Public Health Service Grant GM16735 from the National Institute of General Medical Sciences. The author wishes to express his appreciation to E. J. Capaldi, S. Weinstock, and the reviewers for their helpful comments and criticisms on an earlier draft of this review and again to E. J. Capaldi for supplying prepublication copies of a number of papers.

[2] Requests for reprints should be sent to the author, who is now at Department of Psychology, Emory University, Atlanta, Georgia 30322.

The effect on continuous subjects appears to be greater than the effect on partial subjects. A further complication occurs when the results of a recent study by McCain (1970a) are considered. He gave four groups of rats a 50% reinforcement schedule. The results revealed more resistance to extinction for the group that received a large amount of reinforcement in contrast to one receiving a small amount after 42 acquisition trials. However, this difference disappeared for two comparable groups that had received 138 acquisition trials.

Number of acquisition trials. Most investigators report a positive relation between resistance to extinction after partial reinforcement and total number of training trials (Bacon, 1962; Brooks & Dufort, 1967; Hill & Spear, 1963; Lewis & Cotton, 1959; Wagner, 1961; Wilson, 1964). M. B. Weinstock[3] found that partial subjects after 60 acquisition trials were more resistant to extinction than a partial group receiving 30 trials, while the reverse relationship was found for continuously reinforced groups. However, as reviewed by Sperling (1965a, 1965b), resistance to extinction and number of training trials (under continuously reinforced acquisition) is either inversely, not, or positively related, depending on whether the subjects received a large, moderate, or small amount of reinforcement. This finding obviously complicates the magnitude of the PREE with regard to the number of training trials variable, but, as with amount of reinforcement, the larger effect is apparently on the continuous group rather than the partial group.

Deprivation level. Mikulka and Pavlik (1966) reported an increased magnitude in the PREE as the number of hours of food deprivation was increased. The result appears to arise from an increased resistance to extinction for increased deprivation levels for partially and continuously reinforced subjects, although the increase for continuous subjects was considerably smaller. Ross (1964) found that groups of subjects originally trained under food deprivation (during which time a PRAE was found) then switched to water deprivation (the PRAE disappeared) still demonstrated a PREE under water deprivation, perhaps due to generalization. Finally, Haas, Shessel, Willner, and Rescorla (1970) still found the PREE with groups trained under standard deprivation conditions and extinguished while satiated.

Effects of shifts in reinforcement schedule. Theios (1962) found no difference in resistance to extinction between groups receiving continuously reinforced trials after initial training under partial and a group receiving partial reinforcement throughout. Both groups were more resistant to extinction than continuously reinforced subjects. Rashotte and Surridge (1969) and Harris, Smith, and Weinstock (1962) have also found that the PREE is maintained even though subjects were given a block of continuously reinforced trials after partial training. Sutherland, Mackintosh, and Wolfe (1965) confirmed the latter result and also found that a group receiving partial, followed by continuous, reinforcement was more resistant to extinction than one receiving continuous followed by partial, with subjects receiving partial reinforcement throughout being most resistant. Theios and McGinnis (1967) criticized Sutherland et al. (1965) for failure to adjust for different acquisition levels. When this was done, Theios and McGinnis (1967) reported that although both groups were more resistant to extinction than continuously reinforced subjects, the continuous-shifted-to-partial group was more resistant than the partial-shifted-to-continuous group. However, Leung and Jensen (1968) found that subjects switched from a partial to a continuous reinforcement schedule were more resistant to extinction than both a continuously reinforced group and groups switched from continuous to partial, with the latter groups not differing from the continuous subjects. Thus, although conflicts exist in the literature, a PREE is obtained whether continuously reinforced trials are given before or after a partial schedule. However, the order of schedule presentation is apparently an important variable regarding the magnitude of the obtained extinction effects.

Reextinction. Some investigators view successive reacquisitions and reextinctions as a special case of partial reinforcement (Jensen & Cotton, 1960). Generally, reextinction is slower than original extinction (Cotton, Hill, & Clayton, 1967; Jensen & Cotton, 1960), and sometimes faster asymptotic speeds are obtained (Cotton et al., 1967). However, Capaldi, Leonard, and Ksir (1968) found no change or a decrease in extinction rate for spaced trials and an increase for massed trial groups for successive extinctions.

---

[3] Weinstock, M. B. A factorial study of some variables affecting resistance to extinction under partial reinforcement with spaced trials. Unpublished doctoral dissertation, Indiana University, 1957.

Effects of initial nonreinforcement trials. A number of investigators have shown that subjects given nonreinforced trials prior to 100% reinforced acquisition are more resistant to extinction than 100% reinforced groups, a PREE-like effect (Robbins, Chait, & Weinstock, 1968; Spear & Spitzner, 1967; Spear, Hill, & O'Sullivan, 1965). Recently, Capaldi and Waters (1970) found that a group receiving 5 nonreinforced trials followed by 5 reinforced trials was more resistant to extinction than continuous groups receiving 5 or 10 acquisition trials. However, this effect is usually smaller than the PREE so that these subjects are generally less resistant than partial subjects.

Studies using small numbers of acquisition trials. Recently, there have been a large number of studies in which relatively small numbers of acquisition trials have been followed by extinction trials. Most of these studies have allowed subjects to explore the apparatus or one similar prior to training and have used multiple numbers of food pellets or wet mash as reinforcement. These studies have all shown a PREE (see McCain, 1966, for prior references; McCain & Brown, 1967; Padilla, 1967). In addition, some investigators have given the subjects some small number of continuously reinforced trials followed by a partial schedule (of a few trials) and have also shown a PREE (Bowen & McCain, 1967; McCain, 1966). When prior exploration of the apparatus is eliminated, the PREE has been found by some (Amsel, Hug, & Surridge, 1968; Capaldi & Deutsch, 1967; Capaldi, Lanier, & Godbout, 1968; Capaldi & Waters, 1970), but not by Surridge, Rashotte, and Amsel (1967). This discrepancy may be because Surridge et al. (1967) used a 90-second detention period on both reinforced and nonreinforced trials, while the other studies have used considerably smaller detention periods. Further, with regard to exploration, Godbout, Ziff, and Capaldi (1968) gave groups of rats nonreinforced or reinforced placements in the goal box or a neutral box prior to training and found the PREE and no effect due to the various exposure conditions. When a single pellet has been given as the reinforcement, along with a small number of acquisition trials, the PREE has been found with a 15-second intertrial interval (McCain, 1968) and a 15-minute intertrial interval (McCain, 1969). In contrast, Amsel et al. (1968) failed to find a PREE with trials spaced at one per day for their single-pellet reward group. Thus, a PREE is usually found even when only one or two nonreinforced trials are given during acquisition, although the intertrial interval appears to be a major factor in these studies.

Nonreinforced goal box confinement duration. Hulse (1958) found that a change in nonreinforced goal box confinement duration from acquisition to extinction reduced resistance to extinction; otherwise, no effect due to confinement was found. The former result has been replicated a number of times (Capaldi, 1966b; Capaldi & Poynor, 1965; Stanley & Clayton, 1955), although with regard to confinement time per se, some investigators report an inverse relation between nonreinforced confinement time and resistance to extinction (Capaldi, 1966b; Capaldi & Poynor, 1965). M. B. Weinstock (see Footnote 3) found greater resistance to extinction for a 10-second goal box confinement time in extinction in contrast to a 30-second, and no effect due to nonreinforced confinement time during acquisition. The former effect was larger for the continuous groups than the partial groups.

Goal box size. Both Kirkpatrick, Pavlik, and Reynolds (1964) and Bradley and Wong (1969) have found an increased PREE with a large, in contrast to a small, goal box. This effect seems to be largely attributable to increased resistance to extinction for partials for the larger goal box, with little or no effect being observed on continuously reinforced subjects. Support for this suggestion can be found in the failure to find a difference in resistance to extinction between groups of rats receiving continuous reinforcement in narrow or wide tube alleys and goal boxes (Robbins, 1966). Paul (1969), however, failed to find an effect due to goal box size; it should be noted that she used a shuttle alley and also failed to find the usual PREE.

Goal box cues and activity. Somewhat related to the goal-box-size studies is the Freides (1957) study, where the color of the goal box was manipulated on reinforced and nonreinforced trials during acquisition and extinction. The PREE was obtained for all partial groups although it was larger for groups run to (*a*) a different goal box in extinction than acquisition, (*b*) the negative goal box (from acquisition), (*c*) the same goal box as in acquisition than for a group run to the positive goal box from acquisition.

Pilley and Leeming (1966) tilted their apparatus after the subject entered the goal box during extinction and reported obtaining the PREE with a trial-to-criterion measure, but not with a speed measure. Wilson (1964) fed one group of partially reinforced subjects on the floor of the goal box on some trials and on a shelf in the goal box on others, while two other groups were fed on either the floor or the shelf on their reinforced trials. No differences were found in resistance to extinction. Similarly, Amsel, Surridge, and Hug (1969) found that although more goal approach movements were made during acquisition to a 1-pellet, in contrast to a 10-pellet, reward for partially reinforced subjects, no differences in goal approach movements were found between continuous and partial subjects during extinction, although the conventional PREE (on speed) was found. When activity measures are taken, such as on an open field, partial subjects are generally more active than continuous subjects during extinction (Gallup & Hare, 1969). These studies generally point up differences in goal box activity due to reinforcement schedule.

Measurement of goal box activity. A number of studies, often prompted by interference notions of the PREE, have attempted to measure interfering and competing responses made by the subject in the goal box, and, at times, the alleyway. In summary, partially reinforced subjects make fewer reversals and halts than continuously reinforced ones (Jones & Bridges, 1966, 1967; McCoy & Marx, 1965). Also, a high negative correlation apparently exists between speed and number of halts and reversals (Jones & Bridges, 1966, 1967). Partially reinforced subjects also retrace less than continuous subjects do during extinction in the alleyway (Bradley & Wong, 1969; McCoy & Marx, 1965; Mikulka & Pavlik, 1966). In contrast, Harris et al. (1962) failed to find any differences in the variability of the path taken down the alley between partially and continuously reinforced subjects. When subjects were allowed to retrace the entire length of the alley during extinction, Allison (1967) found that partially reinforced subjects retraced slower than continuously reinforced subjects, who, during acquisition, received the same total number of trials, but faster than a group that received half the number of trials but the same total number of reinforced trials. Longstreth (1964) failed to find any differences between partial and continuous subjects in "backing up" behavior in the goal box. Similarly, Robbins et al. (1968) found that partially reinforced subjects took less time to retrace a section of the goal box than continuously reinforced subjects. Thus, although partial subjects make fewer reversals, halts, and, in general, fewer competing or interfering responses than continuous subjects in both the alley and goal box, they appear to execute these responses (or at least the initial one) faster.

Manipulations in the start box. Bacon (1965) found that subjects blocked 2 inches past the start box door on half of their trials were as resistant to extinction as a conventional partial group. Longstreth (1964) and McCain and Power (1966) found that subjects given 50% reinforcement in the start box (and 100% in the goal box) displayed a PREE. These studies demonstrate that manipulations in the start box, which are generally performed in the goal box, yield the conventional PREE. It should be noted that as in most studies, the start boxes, alleys, and goal boxes were the same color.

Drug studies. Studies using drugs have revealed extremely interesting effects. The general design is to train groups of rats under either partial or continuous schedules, each subject being injected with either saline or a drug, prior to each day's run. Groups are then subdivided during extinction with regard to the drug or saline. Generally, a change in the drug state leads to a smaller PREE, although the effect is still found using amobarital (Gray, 1969; Ison & Pennes, 1969) and chlordiazepoxide (Iwahara, Nagamura, & Iwasaki, 1967). Only Gray (1969) reported an elimination of the extinction effect, and then only when subjects were switched from the drug in acquisition to the saline condition in extinction. It should be noted that these drugs generally reduce the subject's overall speed.

Within-subject effect. Amsel and his colleagues, as well as others (see Amsel, 1967; Amsel, Rashotte, & MacKinnon, 1966, for relevant references), have recently examined a within-subject PREE. Specifically, subjects received partial reinforcement in a distinctive alley (e.g., black) and continuous reinforcement in another distinctive alley (e.g., white). These studies generally reveal a failure to find the conventional PREE, perhaps because of generalization of extinction from the 100% reinforced stimulus to the partially reinforced stimulus. In addition, Rudy, Homzie, Cox, Graeber, and Carter (1970) failed to find greater resistance to extinction in the alley where nonreinforced trials were followed by reinforced trials in contrast to the continuously reinforced alley.

"Patterning." Capaldi and his colleagues (Capaldi, 1966a, 1967) have introduced another major set of variables that emphasize sequential effects of reinforcement schedules. Groups of subjects given a single alternating 50% reinforcement schedule have been found to be less resistant to extinction than a double alternation or continuous group (Bloom & Capaldi, 1961; Capaldi & Minkoff, 1967). In contrast, Surridge and Amsel (1966) found no difference in resistance to extinction between single alternating partial or random partial groups, both being more resistant than a continuous group. The random partial group showed slightly more resistance (although statistically not significant) than the alternating group; this difference was attributed to terminal acquisition differences. Perhaps the use of Anderson's (1963) shape-function measure, which adjusts for different terminal acquisition levels, would help resolve the discrepancy. Capaldi and Hart (1962) pointed out that a single alternating reinforcement schedule leads to more resistance, no difference, or less resistance to extinction than a random partial group, when subjects receive a small, moderate, or large number of training trials, respectively.

Other sequential effects. Capaldi (1966a, 1967) and his colleagues have also investigated the effect, upon resistance to extinction, of the number of consecutive nonreinforced trials preceding a reinforced trial during acquisition (N length), the number of different N lengths, and the number of each N length, while holding the percentage of reinforcement constant. Generally, these studies reveal that independent of the percentage of partial reinforcement, resistance to extinction is positively related to these variables (see Capaldi, 1966a, 1967, for relevant references; Spivey, 1967; Spivey & Hess, 1968; Spivey, Hess, & Black, 1968, 1969). These effects are sometimes interpreted in terms of the number of nonreinforced-to-reinforced trial transitions. Despite the fact that number of training trials may interact with these variables, these studies have demonstrated the importance of sequential effects on the PREE (Capaldi, 1966a, 1967, 1970, in press). Recently, Capaldi and Kassover (1970) found no difference in resistance to extinction between groups receiving different numbers of nonreinforced trials during acquisition with N length and number of nonreinforced-to-reinforced trial transitions held constant. However, N length was found to result in greater resistance to extinction despite a constant number of nonreinforced-to-reinforced trial transitions.

Intertrial reinforcements. Another technique introduced by Capaldi and his colleagues to investigate aspects of sequential theories (Capaldi, 1970) has been to give the subject a reinforced goal box placement after a nonreinforced in contrast to after a reinforced trial. The intertrial reinforcement effect can be interpreted as affecting nonreinforced-to-reinforced trial transitions and N length. Generally, intertrial reinforcements after a nonreinforced run in the alley result in less resistance to extinction than a group receiving its intertrial reinforcement after a reinforced run in the alley (see Capaldi, 1967, for relevant references; Capaldi & Wilson, 1968; McCain, 1966, Experiment IV & V; Spivey, Hess, & Aponte, 1968; Spivey, Hess, & Klemic, 1968). Lobb and Runcie (1967) found that intertrial reinforcements were effective in reducing resistance to extinction if given 15 seconds but not 1 hour after the alley trial. However, Overstreet and Campbell (1969) found that they were not effective if given 15 seconds but were if given 30 seconds after the alley trial.

Recently, evidence has been found that the intertrial reinforcement effect disappears after a relatively large number of acquisition trials (Black & Spence, 1965; Spence, Platt, & Matsumoto, 1965). However, Pratt and Halpern (1969) gave 72 acquisition trials under partial reinforcement and found the effect when they equated experiences with intertrial reinforcement during acquisition or extinction. Since Spence et al. (1965) and Black and Spence (1965) gave placements after every nonreinforced trial, it is possible that subjects eventually discriminated them from regular alley runs, and thus the effect disappeared.

Similarly, groups given partially reinforced goal box placements, when put in front of the food cup and as a result did not locomote much, have shown no PREE when contrasted with a continuously reinforced group. However, when the subject is placed at the front of the goal box, so that locomotion is obtained, the PREE is obtained (Trapold & Doren, 1966; Trapold & Holden, 1966). In contrast, McCain (1970b) and Wilton (1967) found a PREE after partially reinforced goal box placements in front of the food cup. In addition, Homzie, Rudy, and Carter (1970) found that with a small number of acquisition trials, intertrial reinforcements eliminated the PREE, but with extended training, a PREE was obtained, that is, the effect of intertrial reinforcements was eliminated. They also reported that intertrial nonreinforcements led to a PREE with a small number of trials but obtained no PREE with extended training. Theios and Polson (1962) found that subjects given partially reinforced runs in the alley were more resistant to extinction than a group receiving continuously reinforced alley runs and partially reinforced placements. These studies suggest that nonreinforced "runs" in the alley and even in the goal box may not be essential for the appearance of the PREE. Perhaps, as suggested by Wilton (1967), exposure to the goal area is essential, and this may constitute "approach responses" to the goal.

ACQUISITION EFFECT

Although almost all investigators report finding the PREE, many have failed to find the PRAE. The difference in the magnitude of the effects is quite large, the PRAE being considerably smaller. The boundary conditions for the acquisition effect appear to be relatively widely spaced trials and a relatively large number of acquisition trials. There are actually three acquisition "effects." Initially, partial subjects run slower than continuous subjects, although ultimately they run faster. This is often referred to as the "cross-over effect" (Goodrich, 1959; Haggard, 1959; Ross, 1964; Spear et al., 1965; Wagner, 1961; Weinstock, 1958). Asymptotically then, partial subjects run faster than continuous subjects. The asymptotic difference, the second effect, is the one that is primarily dealt with here, so that reference to the PRAE will mean faster asymptotic running by partial, in contrast to continuous, subjects. Some investigators have recorded goal box speeds, in addition to start box and alley speeds, and have reported faster asymptotic running speeds by continuous subjects, that is, no PRAE (Goodrich, 1959; Haggard, 1959; McCoy & Marx, 1965; Wagner, 1961). This effect is dealt with separately and is referred to as the "goal box" effect. The basic reference for the acquisition effect is again Weinstock (1958), who found an inverse relation between asymptotic start and alley speeds and the trial block of the crossover, and percentage of reinforced trials.

Amount of reinforcement. A number of investigators have reported faster asymptotic speeds for partial subjects receiving large, in contrast to small, amounts of reinforcement (Brooks & Dufort, 1967; Hill & Wallace, 1967; Hulse, 1958; Mikulka, Lehr, & Pavlik, 1967; Padilla, 1967; Wagner, 1961), although recently both Amsel, Surridge, and Hug (1969) and Roberts (1969) failed to find this using varying numbers of food pellets. Of the above studies, only Amsel, Surridge, and Hug (1969) and Mikulka et al. (1967) found the PRAE. These data suggest that amount of reinforcement has similar effects on continuous and partial subjects, since increasing amount or duration of reinforcement generally leads to faster asymptotic running speed under continuous reinforcement (Robbins, 1969). However, McCain (1970a) recently reported that a partial group given a large, in contrast to a small, amount of reinforcement ran faster after 42 acquisition trials, although this difference disappeared after 138 trials. The disappearance or reduction of the effect of amount of reinforcement on both continuously and partially reinforced subjects, as pointed out by McCain (1970a), requires further investigation, and a definitive statement must await such research.

Deprivation level. Mikulka and Pavlik (1966) reported finding the crossover effect for only an ad lib group with no differences between partial and continuous groups under various deprivation levels. On the other hand, Badia (1965) found the crossover effect under $2\frac{1}{2}$-, $11\frac{1}{2}$-, and $22\frac{1}{2}$-hour food deprivation, and, although partially reinforced subjects ran faster asymptotically under medium and high deprivation conditions, these differences were not statistically significant. It appears that increasing deprivation level results in faster asymptotic running for both partial and continuous subjects, but Badia's (1965) data suggest that the effect on partial subjects levels off faster than that on continuous subjects. Finally, Ross (1964) found that when groups of subjects, after demonstrating a PRAE under food deprivation, were switched to water deprivation, the asymptotic difference disappeared.

Effects of shifts in reinforcement schedules. Groups of rats after partial reinforcement training often show an increase in running speed when switched to a 100% reinforcement schedule (Harris et al., 1962; Hill & Wallace, 1967; Mikulka et al., 1967; Robbins et al., 1968; Sutherland et al., 1965), although others have failed to obtain this effect (Badia, 1965; Rashotte & Surridge, 1969; Surridge, Boehnert, & Amsel, 1966; Theios, 1962). Groups of rats after 100% training, when shifted to a partial reward schedule, usually show a decrease in speed (Mikulka et al., 1967; Sutherland et al., 1965). Badia (1965) found that subjects switched from partial to continuous or continuous to partial converged to one another's asymptotic level. Leung and Jensen (1968) gave groups of rats different percentages of reinforcement followed by continuous reinforcement and found that they all ran faster than subjects continuously reinforced throughout. Similarly, subjects switched from continuous to partial schedules generally showed no effect or a decrease in speed. Thus, subjects switched from continuous to partial generally decline in speed, and the reverse leads to no change or an increase in speed.

Reacquisition. Successive reacquisition and reextinction may be viewed as similar in form to the partial reinforcement situation (Jensen & Cotton, 1960). Generally, reacquisition is faster than original acquisition (Cotton et al., 1967; Harris et al., 1962; Huang & Huang, 1970; Jensen & Cotton, 1960; North & Morton, 1962), and sometimes higher asymptotes are obtained (Cotton et al., 1967; Harris et al., 1962). However, Surridge et al. (1966) failed to find higher reacquisition asymptotes after 80 extinction trials, although Robbins et al. (1968) did. The reasons for these discrepancies are not clear.

Effect of initial nonreinforced trials. A number of investigators have found that groups of rats given nonreinforced trials prior to 100% reinforced acquisition run faster at asymptote than subjects receiving 100% reinforcement throughout (Cotton et al., 1967; Logan, 1960, p. 188; Robbins et al., 1968; Spear & Spitzner, 1967), while others have failed to find this effect (Spear et al., 1965). It appears that relatively widely spaced trials are requisite, or at least helpful, in obtaining this particular effect.

Nonreinforced goal box confinement duration. Hulse (1958) found no effect on acquisition asymptotes due to nonreinforced goal box confinement time. It should be noted that he gave only 25 acquisition trials, and this may account for the failure to obtain the PRAE, so that a definitive statement cannot be made.

Goal box size. Both Bradley and Wong (1969) and Kirkpatrick et al. (1964) failed to find any difference between partial and continuous subjects as a function of goal box size. Goodrich (1959) found the PRAE in a small and large goal box, and inspection of the curves suggests a bigger effect due to the large goal box. In addition, the large-goal-box study yielded no differences in goal speed between partials and continuous subjects, while the small goal box yielded a significant "goal box effect."

Goal box cues and activity. Wilson (1964) found little difference between a partial group fed either on the floor or on a shelf in the goal box on different trials and a partial group fed in only one location through acquisition, with a slightly greater speed displayed by the latter group. Tacker and Way (1968) replaced the goal box with a large open field and found that partial subjects were more active after nonreinforced in contrast to reinforced trials, although the differences diminished during training, but not to zero. Gallup and Hare (1969) reported that partial subjects were found to be more active (in an open field) after both reinforced and nonreinforced trials than continuously reinforced subjects. Goodrich (1959) found that partials spent more time at the food cup than continuous subjects on reinforced trials. These studies, in general, reveal that partial subjects are more active in the goal box than continuous subjects.

Measurement of goal box activity. McCoy and Marx (1965) found that partial subjects made more reversals and halts in the alley early in training than continuous subjects, although at asymptote both groups were making almost no competing responses. A similar result has been reported by Jones and Bridges (1966), although the crossover effect was not statistically significant in a study using satiated rats (Jones & Bridges, 1967). Goodrich (1959) found that partials stopped more in the alley early in training but, later in training, converged to the lower level of continuous subjects. On the other hand, Mikulka and Pavlik (1966) found that partials made slightly more pauses and reversals in the runway than continuous subjects under various deprivation levels, except for a satiated condition, where the reverse was found. Harris et al. (1962) reported no change in the variability of the path taken by partial subjects on nonreinforced trials during acquisition. McCoy and Marx (1965) also measured pauses and reversals in the goal box and found that partials made more than continuous subjects throughout acquisition. Partial subjects have also been found to retrace faster than continuous subjects on reinforced trials (Allison, 1967), and there has also been a failure to find an increase in retrace time during nonreinforced acquisition trials (Allison, 1967; Robbins et al., 1968). In addition, Longstreth (1964) found an increase in competing behavior on nonreinforced trials during acquisition. In summary, it appears that partial subjects perform less competing activity in the runway, more in the goal box, and show little change in time to perform the initial interfering response in the goal box during the course of acquisition. It should be noted that of the above studies, only McCoy and Marx (1965), Harris et al. (1962), and Robbins et al. (1968) found a PRAE, and that McCoy and Marx (1965) also reported finding the "goal box" effect.

Manipulation in the start box. Bacon (1965) found no differences in asymptotic acquisition speeds between subjects blocked from the alley on half of their trials, a conventional partial group, and a 100% reinforced group. When subjects are fed on half of their trials in the start box (and 100% of the time in the goal box), they are found to run faster after nonreinforced, in contrast to reinforced (start box), trials (Longstreth, 1964; McCain & Power, 1966). The faster running after nonreinforcement is similar to an effect (the "frustration" effect) often found by Amsel and others in the "double-alley" (Amsel, 1958, 1962). They reported faster running in the second alley after nonreinforcement in the goal box of the first alley, in contrast to running after reinforcement in the first goal box.

Drug studies. Generally, the PRAE is eliminated with amobarbital (Gray, 1969; Wagner, 1963), chlordiazepoxide (Iwahara, Iwasaki, Nagamura, & Masiryama, 1966), and ethyl alcohol (Nelson & Wollen, 1965), with controls (both partial and continuous subjects) running faster than experimental subjects.

Within-subject effect. Using the within-subject procedure previously described, Amsel and his colleagues and others (see Amsel, 1967; Amsel et al., 1966, for relevant references) have generally found the conventional PRAE in the start box and alley, and the reversal in the goal box.

"Patterning." Capaldi and his colleagues gave subjects a single alternating 50% reinforcement schedule and found a "patterning effect," that is, faster running following nonreinforced trials in contrast to trials following reinforcement, that is, faster running on reinforced in contrast to nonreinforced trials (Bloom, 1967; Bloom & Capaldi, 1961; Bowen & Strickert, 1966; see Capaldi, 1967, for other relevant references; Ison & Pinckney, 1968; McCain, 1966; Experiment IV). The faster running after nonreinforcement is reminiscent of the "frustration" effect referred to earlier, in the double alley (Amsel, 1958, 1962). With a double-alternation schedule, no patterning effect was found by Bloom and Capaldi (1961), although Yelen (1967) did find a patterning effect with both large and small rewards. Generally, when the discrimination develops, subjects slow down on nonreinforced, in contrast to reinforced, trials. However, conflicts with regard to the patterning effect exist in the literature. Amsel and his colleagues have suggested that the experimenter may inadvertently provide the subject with cues with regard to trial type (e.g., cleaning out the goal box only after reinforced trials or differential odor cues) and have provided evidence in this regard (Surridge & Amsel, 1965a). Recently, some studies in which trials were spaced at one per day have failed to detect a patterning effect (Surridge & Amsel, 1965b, 1966, 1968). However, others have found patterning with one trial per day (Capaldi & Lynch, 1966; Capaldi & Spivey, 1964a, 1964b). Capaldi and Spivey (1965) stated that the discrepancy may be related to nonreinforced trial goal box confinement time. They cited the data of Burt and Wike (1963), which showed that nonreinforced trial speeds are slower and patterning appears earlier the longer the nonreinforced trial confinement. In addition, they provided evidence that at one trial per day, patterning results from 120-second confinement times but not from 30-second confinement time (Capaldi & Lynch, 1966). Bloom (1967), using highly massed trials, found that large reinforcements and large nonreinforced confinements produced the largest patterning effect. However, with the one trial per day procedure, Surridge and Amsel (1968) failed to find patterning using three groups that had reinforced-nonreinforced durations of 30 seconds–30 seconds, 30 seconds–120 seconds, and 120 seconds–30 seconds, respectively. They suggested that perhaps reward magnitude is an important variable in obtaining the effect, although pertinent data are not available. Amsel et al. (1968) did not dispute the findings with highly massed trials, and, in fact, data from their laboratory revealed patterning with a 20-second intertrial interval but not with a 12-minute interval (Surridge, 1967). They have failed to obtain patterning at one trial per day and have pointed out many of the possible sources of differential cues the experimenter may inadvertently provide as a basis for the effect (Surridge & Amsel, 1965b, 1968). Bloom and Malone (1968), with a 1-hour interval, and Harris and Thomas (1966), with highly massed trials, reported finding the patterning effect when care was taken to eliminate possible sources of differential cues. The discrepancies are not yet resolved, but some summary statements may be advanced. Under highly massed trials a patterning effect appears. However, with spaced trials the results are unclear since relatively long nonreinforced *and* reinforced goal box confinement times in some cases lead to patterning (Bloom & Malone, 1968; Capaldi & Lynch, 1966; Capaldi & Spivey, 1964a, 1964b), while in others patterning is not found (Surridge, 1967; Surridge & Amsel, 1965a, 1966, 1968). Obviously, additional research is required and, as pointed out by Surridge and Amsel (1968): "The reason for this discrepancy, apparent or real, is still not entirely obvious [p. 108]."

Recently, studies have been reported (e.g., Ludvigson & Sytsma, 1967; McHose, 1967, McHose, Jacoby, & Meyer, 1967) that suggest that rats may exude differential odors when nonreinforced, in contrast to those exuded when reinforced, and that rats can use odor trails (of the previous subject) as a differential cue if subjects are run in squads. Based on these studies, Amsel, Hug, and Surridge (1969) found that at one trial per day, they were able to obtain a patterning effect, eliminate it, and reinstate it, depending upon whether the reinforcement or nonreinforcement of the first rat run each day was determined or random. They interpreted their results as evidence that odors may be used as differential cues when they are systematically correlated with reinforcement and/or nonreinforcement. However, their data are in conflict with those of Capaldi and Lynch (1966), who took care to control differential odor cues and found a patterning effect.

Other sequential effects and intertrial reinforcements. The studies investigating the effects of nonreinforced-to-reinforced trial transitions and N length on resistance to extinction, either through schedule manipulation or intertrial reinforcements, have generally failed to reveal any terminal acquisition differences. Although effects on extinction have been found, most studies fail to detect terminal acquisition differences.

The goal box effect. It has been pointed out (Robbins et al., 1968) that in those studies that have found the PRAE in the start box and alley, but not in the goal box, a potential discriminative cue on nonreinforced as opposed to reinforced trials may have been operating. In these studies, on nonreinforced trials either an empty food cup or no food cup was present, and the goal box measure was taken to a distance close to the food cup area (Badia, 1965; Goodrich, 1959; Haggard, 1959; McCoy & Marx, 1965; Wagner, 1961). Although Goodrich (1959) reported no differences in speed on reinforced and nonreinforced trials, it seems that an experimental control when available, rather than a statistical comparison, is generally more desirable. When care is taken to remove this source of possible discriminative stimulation for the presence or absence of reinforcement, the standard PRAE is obtained under conditions of satiation (Dachowski & Dunlap, 1969) as well as under standard deprivation conditions (Robbins et al., 1968).

SOME THEORETICAL POSITIONS

Lewis (1960) summarized the major theoretical positions popular at that time. In this review, we evaluate three intertrial or nonsequential theories (Capaldi, 1970) and a sequential theory (Capaldi, 1966a, 1967, 1970, in press). Since the extinction effect is most robust, theoretical analyses of this effect are evaluated initially, and if the theory is still viable, then the acquisition effect will be dealt with.

Cognitive dissonance. Lawrence and Festinger (1962) have applied dissonance theory to the behavior of the rat in the alleyway. Nonreinforcement is assumed to produce dissonance because the organism has expended effort to reach the goal box (and presumably "expects" reward) and receives no reinforcement. Since, in the usual partial schedule, reinforcements and nonreinforcements are intermixed in some random order, the subject is assumed to develop some extra preference either for the activity or for the goal box itself. Thus, in extinction, partial subjects have this extra "attraction," while continuous subjects do not, and therefore the PREE is expected. The patterning effect on extinction, and the other sequential effect studies, reveal that it is the trial sequences, and not simply total number of nonreinforced trials as suggested by Lawrence and Festinger (1962), that are related to resistance to extinction. In addition, studies demonstrating increased resistance to extinction after initial nonreinforced trials may be damaging to this view since opportunities for "extra attractiveness" to be built up are not available unless it is assumed that nonreinforced trials themselves lead to "extra attractiveness." Finally, studies using a small number of acquisition trials, which show a PREE, are damaging for similar reasons. Thus, there is considerable evidence that provides difficulty for the dissonance view of the PREE.

"Frustration." According to the Amsel and Spence formulation (Amsel, 1958, 1962; Spence, 1956, 1960), nonreinforcements lead to the development of "frustration" only when they are preceded by reinforced trials.

Thus, frustration develops due to the subject not being reinforced after an expectancy of reward has been built up. Spence (1956, 1960) has further argued that partial subjects should run faster in the alley because they are running under a "frustrative" drive, but that continuous subjects should run faster in the goal box because they have developed higher levels of "incentive motivation." Studies using a small number of acquisition trials, revealing a PREE, are particularly troublesome for frustration theory since not many trials are given, and it would appear that there is not enough opportunity for "frustration" to be built up. Some theorists have suggested that there are two partial effects, a small trials effect, and the more usual one (Amsel et al., 1968). The latter is presumably based on frustration concepts. In fact, Amsel et al. (1968) modified frustration theory so that one reinforced trial was enough to provide frustration by resorting to a buildup of reward expectancy through the use of multiple-pellet in contrast to single-pellet rewards. However, studies giving initial nonreinforced trials, revealing a "partial extinction-like" effect have shown, as stated by Capaldi, Ziff, and Godbout (1970), that "If reward is expected here, the source cannot be a prior rewarded trial [p. 62]." Brooks (1969) recently suggested that the small trials PREE may result from differences in the amount of "primary *unconditioned* frustration" between continuous and partial groups. This presumably results from continuous subjects receiving more reinforced trials than partials, since, in these studies, the total number of trials is usually equated. A similar interpretation has also been suggested by Amsel, Surridge, and Hug (1969). However, Capaldi and Waters (1970) found that even when the number of reinforced trials was equated (five), groups receiving five nonreinforced trials either prior to their reinforced trials or following two reinforced trials were more resistant to extinction than the continuously reinforced (five-trial) group. Thus, without additional explanatory concepts, evidence against a frustrative view of the extinction effect exists. Attention should also be brought to the "goal box" effect or "Goodrich-Wagner" effect, since the reversal of the PRAE in the goal box during acquisition is often cited as support for the frustration position. However, as shown by Robbins et al. (1968) and Dachowski and Dunlap (1969), when discriminative stimulation due to presence or absence of the reward is eliminated, the reversal of the PRAE in the goal box disappears. These results, in addition to the initial nonreinforced trial studies and the small number of acquisition trial studies, provide a body of data that constitutes considerable difficulty for frustration theory.

"Competing responses." The most current version of competing response theory is that of Weinstock (Robbins et al., 1968; Weinstock, 1954, 1958). According to this view, nonreinforced trials provide an opportunity for nonfunctional components of the running response to drop out, that is, to habituate and thereby enable partially reinforced animals to achieve a higher acquisition asymptote than continuously reinforced subjects and also display more resistance to extinction. Despite the fact that a number of predicted effects on running speed have been confirmed (e.g., reacquisition, effect of initial nonreinforced trials), there has been a general failure to detect the elimination or habituation of competing responses (e.g., Harris et al., 1962; McCoy & Marx, 1965; Robbins et al., 1968). In addition, studies using small numbers of acquisition trials are troublesome unless one is willing to assume that habituation occurs at a very high rate. Thus, although speed effects are found, problems arise, since the basic mechanism has not received independent empirical support. As a result, when competing response theory is specified, as suggested by Marx (1963), it leads to failure to validate empirically the major construct.

Sequential theory. Capaldi (1966a, 1967, 1970, in press) has recently revived the old aftereffects theory by invoking a memory mechanism. Specifically, it is assumed that stimuli arising from nonreinforcement are maintained through a memory mechanism, and when the subject is eventually reinforced, these stimuli become conditioned to the running response. The patterning studies, intertrial reinforcement studies, and the other sequential effects studies have all been taken as providing strong support for this view. Additional support for the sequential theory comes from a series of studies investigating sequence effects on resistance to extinction after training with varying amounts of partial reward. Generally, large amounts of reinforcement lead to greater resistance to extinction under a partial schedule than small amounts. However, it has been shown that the critical variable is the number of nonreinforcements preceding a large amount of reinforcement; this condition leads to greater resistance to extinction than groups given the same number of small and large rewards and nonreinforcements, except in a different order (Capaldi, 1970; Capaldi & Lynch, 1968; Capaldi & Minkoff, 1969; Leonard, 1969). Some have suggested that with massed trials, sequential variables are crucial, but with relatively spaced trials either a frustrative view (Amsel, 1967) or a contrast view (Gonzalez & Bitterman, 1969) is appropriate. However, some of the partial reward-sequential studies referred to above were at one trial per day. Thus, the sequential view appears to be most viable with regard to the extinction effect. As pointed out by Capaldi (1970), the theory is an extinction theory. He made the argument that extinction provides a better measure regarding what is learned in acquisition than "relatively insensitive acquisition measures." He points out that subjects learn different things during acquisition depending upon sequences, and behavior during extinction is the best indicator of what was learned in acquisition.

To summarize up to this point, difficulty exists for the cognitive dissonance, frustration, and competing response views, primarily because of studies revealing a PREE under conditions in which sufficient time or trials does not appear to have been available for the development of either "extra attractiveness," "frustration," or "habituation." Further, the sequential theory proposed by Capaldi appears to capture the major aspects of the extinction effect. Application of this view to acquisition would appear to predict presence or absence of a PRAE depending upon whether one or both processes, that is, of stimulus modification arising from nonreinforcement and conditioning that resulted from reinforcement, had reached asymptote. It would appear, however, that if both processes had reached their respective asymptotes, subjects should run faster after a nonreinforced trial, in contrast to after a reinforced trial. Aside from results of "patterning" studies, there appears to be no evidence for such a difference (e.g., Goodrich, 1959; Weinstock, 1958). Thus, it appears that all three nonsequential theories evaluated, without additional concepts or modifications, as well as Capaldi's sequential theory, cannot account for the entire range of data. It may be the case, as pointed out by Weinstock (1970), that a position incorporating elements of each may be most viable. However, this would undoubtedly be somewhat cumbersome. As a result, for the theorist desiring parsimony, along with explanatory validity, the evidence reviewed here appears to suggest that the sequential approach may be a viable approach and worth considerable, theoretical, as well as experimental, exploration.

The basic mechanisms of sequential theory may be cast in the following framework: Nonreinforcement has specific stimulus consequences that persist through some long-term memory or storage mechanism. Successive nonreinforcements lead to a modification of the stored stimulus consequences, presumably to some asymptotic strength. Reinforcement leads to the conditioning of stimuli present on that trial to the running response; the stimulus source is not only external alley cues but also the representation (in memory) of stimuli arising from nonreinforcement. Thus, nonreinforcement during extinction leads to stimuli that have been conditioned to running during acquisition for those groups given nonreinforced trials followed by reinforcements. In addition, longer nonreinforced trial runs followed by reinforcement, as well as increasing numbers of nonreinforced-to-reinforced trial transitions, should lead to greater stimulus modification (arising from nonreinforcement) and thus a greater habit strength with regard to the running response.

Some concluding remarks regarding the effects of the variables reviewed in this study are in order. It is both interesting and disturbing that some variables have different effects depending upon whether the subject is partially or continuously reinforced. Since most studies have used a single partial schedule (50% reinforcement), the generality of the former interaction is not known. It would seem unlikely that basically different mechanisms operate depending on reinforcement schedule. For example, how different is a 90% reinforcement schedule from a 100% schedule? Thus, a further task is to unravel this apparently dichotomous situation. Further investigations and extensions of sequential theory may shed light on this problem.

REFERENCES

Allison, J. Time spent in the goal box: Effects of frequency and schedule of reinforcement. *Psychonomic Science,* 1967, **7**, 165–166.

Amsel, A. The role of frustrative nonreward in noncontinuous reward situation. *Psychological Bulletin,* 1958, **55**, 102–119.

Amsel, A. Frustrative nonreward in partial reinforcement and discrimination learning: Some recent history and a theoretical extension. *Psychological Review,* 1962, **69**, 306–328.

Amsel, A. Partial reinforcement effect on vigor and persistence: Advances in frustration theory derived from a variety of within-subjects experiments. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory*. Vol. 1. New York: Academic Press, 1967.

Amsel, A., Hug, J. J., & Surridge, C. T. Number of food pellets, goal approaches, and the partial reinforcement effect after minimal acquisition. *Journal of Experimental Psychology,* 1968, **77**, 530–534.

Amsel, A., Hug, J. J., & Surridge, C. T. Subject-to-subject trial sequence, odor trails, and patterning at 24-h ITI. *Psychonomic Science,* 1969, **15**, 119–120.

Amsel, A., Rashotte, M. E., & MacKinnon, J. R. Partial reinforcement effects within subject and between subjects. *Psychological Monographs,* 1966, **80**(20, Whole No. 628).

Amsel, A., Surridge, C. T., & Hug, J. J. Number of food pellets and the partial reinforcement extinction effect after extended acquisition. *Journal of Experimental Psychology,* 1969, **82**, 578–580.

Anderson, N. H. Comparisons of different populations: Resistance to extinction and transfer. *Psychological Review,* 1963, **70**, 162–179.

Bacon, W. E. Partial reinforcement extinction effects following different amounts of training. *Journal of Comparative and Physiological Psychology,* 1962, **55**, 998–1003.

Bacon, W. E. Resistance to extinction following blocking of the instrumental response during acquisition. *Journal of Experimental Psychology,* 1965, **69**, 515–521.

Badia, P. Effects of drive, reinforcement schedule, and change of schedule on performance. *Journal of Experimental Psychology,* 1965, **69**, 292–297.

Black, R. W., & Spence, K. W. Effects of intertrial reinforcement on resistance to extinction following extended training. *Journal of Experimental Psychology,* 1965, **70**, 559–563.

Bloom, J. M. Early acquisition responding on trials following different rewards and nonrewards. *Psychonomic Science,* 1967, **7**, 37–38.

Bloom, J. M., & Capaldi, E. J. The behavior of rats in relation to complex patterns of partial reinforcement. *Journal of Comparative and Physiological Psychology,* 1961, **54**, 261–265.

Bloom, J. M., & Malone, P. Single alternation patterning without a trace for blame. *Psychonomic Science,* 1968, **11**, 335–336.

Bowen, J., & McCain, G. Occurrence of the partial reinforcement extinction effect after only one NRNR sequence of trials. *Psychonomic Science,* 1967, **9**, 15–16.

Bowen, J., & Strickert, D. Discrimination learning as a function of internal sitmuli. *Psychonomic Science,* 1966, **5**, 297–298.

Bradley, H. W., & Wong, R. Extinction as a function of size of goal box, reinforcement schedule, and competing responses. *Psychonomic Science,* 1969, **17**, 189–190.

Brooks, C. I. Frustration to nonreward following limited reward experience. *Journal of Experimental Psychology,* 1969, **81**, 403–405.

Brooks, C. I., & Dufort, R. H. Resistance to extinction as a function of amount of reinforcement and reinforcement schedule. *Psychonomic Science,* 1967, **9**, 165–166.

Burt, D. H., & Wike, E. L. Effects of alternating partial reinforcement and alternating delay of reinforcement on a runway response. *Psychological Reports,* 1963, **13**, 439–442.

Capaldi, E. J. Partial reinforcement: A hypothesis of sequential effects. *Psychological Review,* 1966, **73**, 459–477. (a)

Capaldi, E. J. Stimulus specificity: Nonreward. *Journal of Experimental Psychology,* 1966, **72**, 410–414. (b)

Capaldi, E. J. A sequential hypothesis of instrumental learning. In K. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation.* New York: Academic Press, 1967.

Capaldi, E. J. An analysis of the role of reward and reward magnitude in instrumental learning. In J. H. Reynierse (Ed.), *Current issues in animal learning.* Lincoln: University of Nebraska Press, 1970.

Capaldi, E. J. Memory and learning. A sequential viewpoint. In W. Honig & N. Macintosh (Eds), *Dalhousie symposium on animal memory.* Halifax: Dalhousie University Press, in press.

Capaldi, E. J., & Deutsch, E. A. Effects of severely limiting acquisition training and pretraining on the partial reinforcement effect. *Psychonomic Science,* 1967, **9**, 171–172.

Capaldi, E. J., & Hart, D. Influence of a small number of partial reinforcement training trials on resistance to extinction. *Journal of Experimental Psychology,* 1962, **64**, 166–171.

Capaldi, E. J., & Kassover, K. Sequence, number of nonrewards, anticipation, and intertrial interval in extinction. *Journal of Experimental Psychology,* 1970, **84**, 470–476.

Capaldi, E. J., Lanier, A. T., & Godbout, R. C. Reward schedule effects following severely limited acquisition training. *Journal of Experimental Psychology,* 1968, **78**, 521–524.

Capaldi, E. J., Leonard, D. W., & Ksir, C. A reexamination of extinction rate in successive acquisitions and extinctions. *Journal of Comparative and Physiological Psychology,* 1968, **66**, 128–132.

Capaldi, E. J., & Lynch, D. Patterning at 24-hour ITI: Resolution of a discrepancy more apparent than real. *Psychonomic Science,* 1966, **6**, 229–230.

Capaldi, E. J., & Lynch, A. D. Magnitude of partial reward and resistance to extinction: Effect of N-R transitions. *Journal of Comparative and Physiological Psychology,* 1968, **65**, 179–181.

Capaldi, E. J., & Minkoff, R. Reward schedule effects at a relatively long intertrial interval. *Psychonomic Science,* 1967, **9**, 169–170.

Capaldi, E. J., & Minkoff, R. Influence of order of occurrence of nonreward and large and small reward on acquisition and extinction. *Journal of Experimental Psychology,* 1969, **81**, 156–160.

Capaldi, E. J., & Poynor, H. B. Nonreward confinement duration: The complement of reward magnitude. *Psychonomic Science,* 1965, **3**, 515–516.

Capaldi, E. J., & Spivey, J. E. Intertrial reinforcement and aftereffects at 24-hour intervals. *Psychonomic Science,* 1964, **1**, 181–182. (a)

Capaldi, E. J., & Spivey, J. E. Stimulus consequences of reinforcement: Stimulus traces or memory. *Psychonomic Science,* 1964, **1**, 403–404. (b)

Capaldi, E. J., & Spivey, J. E. Comment. *Psychonomic Science,* 1965, **3**, 110.

Capaldi, E. J., & Waters, R. W. Conditioning and nonconditioning interpretations of small-trial phenomena. *Journal of Experimental Psychology,* 1970, **84**, 518–522.

Capaldi, E. J., & Wilson, N. B. Intertrial reinforcement: A test of several hypotheses. *Psychonomic Science,* 1968, **13**, 169–170.

Capaldi, E. J., Ziff, D. R., & Godbout, R. C. Extinction and the necessity or non-necessity of anticipating reward on nonrewarded trials. *Psychonomic Science,* 1970, **18**, 61–62.

Cotton, J. W., Hill, W. F., & Clayton, K. N. The effect of varying numbers of rewarded and nonrewarded trials on subsequent performance in a straight runway. *Psychonomic Science,* 1967, **9**, 489–490.

Dachowski, L., & Dunlap, W. P. Frustrative nonreward effects in acquisition under zero hours deprivation. *Psychonomic Science,* 1969, **14**, 113–114.

Freides, D. Goal-box cues and pattern of reinforcement. *Journal of Experimental Psychology,* 1957, **53**, 361–371.

Gallup, G. G., Jr., & Hare, G. K. Activity following partially reinforced trials: Evidence for a residual frustration effect due to conditioned frustration. *Psychonomic Science,* 1969, **16**, 41–42.

Godbout, R. C., Ziff, D. R., & Capaldi, E. J. Effect of several reward exposure procedures on the small trial PRE. *Psychonomic Science,* 1968, **13**, 153–154.

Gonzalez, R. C., & Bitterman, M. E. Spaced-trials partial reinforcement effects as a function of contrast. *Journal of Comparative and Physiological Psychology,* 1969, **67**, 94–103.

Gonzalez, R. C., Roberts, W. A., & Bitterman, M. E. Resistance to extinction in normal and extensively decorticated rats as a function of amount and percentage of reinforcement. *American Journal of Psychology,* 1966, **79**, 542–550.

Goodrich, K. P. Performance in different segments of an instrumental response chain as a function of reinforcement schedule. *Journal of Experimental Psychology,* 1959, **57**, 57–63.

Gray, J. A. Sodium amobarbital and effects of frustrative nonreward. *Journal of Comparative and Physiological Psychology,* 1969, **69**, 55–64.

Haas, R. B., Shessel, F. M., Willner, H. S., & Rescorla, R. A. The effect of satiation following partial reinforcement. *Psychonomic Science,* 1970, **18**, 296–297.

Haggard, D. F. Acquisition of a simple running response as a function of partial and continuous schedules of reinforcement. *Psychological Record,* 1959, **9**, 11–18.

Harris, S. J., Smith, M. G., & Weinstock, S. Effects of nonreinforcement on subsequent reinforced running behavior. *Journal of Experimental Psychology,* 1962, **64**, 388–392.

Harris, J. H., & Thomas, G. J. Learning single alternation of running speeds in a runway without handling between trials. *Psychonomic Science,* 1966, **6**, 329–330.

Hill, W. F., & Spear, N. E. Extinction in a runway as a function of acquisition level and reinforcement percentage. *Journal of Experimental Psychology,* 1963, **65**, 495–500.

Hill, W. F., & Wallace, W. P. Effects of magnitude and percentage reward on subsequent patterns of runway speed. *Journal of Experimental Psychology,* 1967, **73**, 544–548.

Homzie, M. J., Rudy, J. W., & Carter, E. N. Runway performance in rats as a function of goal-box placements and goal-event sequence. *Journal of Comparative and Physiological Psychology,* 1970, **71**, 283–291.

Huang, I., & Huang, H. The effect of initial reward on the rate of reacquisition. *Psychonomic Science,* 1970, **18**, 39–40.

Hulse, S. H., Jr. Amount and percentage of reinforcement and duration of goal confinement in conditioning and extinction. *Journal of Experimental Psychology,* 1958, **56**, 48–57.

Ison, J. R., & Pennes, E. S. The joint effect of amobarbital sodium and partial reinforcement on runway acquisition and extinction. *Journal of Comparative and Physiological Psychology,* 1969, **68**, 215–219.

Ison, J. R., & Pinckney, L. A. Facilitative effects of reward and decremental effects of non-reward in the initial trials of partial reinforcement. *Psychonomic Science,* 1968, **12**, 37–38.

Iwahara, S., Iwasaki, T., Nagamura, N., & Masiryama, E. Effect of chlordiadepoxide upon partially reinforced behavior of rats in the straight runway. *Japanese Psychological Research,* 1966, **8**, 131–135.

Iwahara, S., Nagamura, N., & Iwasaki, T. Effect of chlordiadepoxide upon experimental extinction in the straight runway as a function of partial reinforcement in the rat. *Japanese Psychological Research,* 1967, **9**, 128–134.

Jenkins, W. O., & Stanley, J. C. Partial reinforcement: A review and critique. *Psychological Bulletin,* 1950, **47**, 193–234.

Jensen, G. D., & Cotton, J. W. Successive acquisition and extinctions as related to percentage of reinforcement. *Journal of Experimental Psychology,* 1960, **60**, 41–49.

Jones, E. C., & Bridges, C. C., Jr. Competing responses and the partial reinforcement effect. *Psychonomic Science,* 1966, **6**, 483–484.

Jones, E. C., & Bridges, C. C., Jr. The partial reinforcement effect and competing behavior in a latent learning situation. *Psychonomic Science,* 1967, **7**, 119–120.

Kirkpatrick, D. R., Pavlik, W. B., & Reynolds, W. F. Partial reinforcement extinction effect as a function of the size of the goal box. *Journal of Experimental Psychology,* 1964, **68**, 515–516.

Lawrence, D. H., & Festinger, L. *Deterrents and reinforcement.* Stanford: Stanford University Press, 1962.

Leonard, D. W. Amount and sequence of reward in partial and continuous reinforcement. *Journal of Comparative and Physiological Psychology,* 1969, **67**, 204–211.

Leung, C. M., & Jensen, G. D. Shifts in percentage of reinforcement viewed as changes in incentive. *Journal of Experimental Psychology,* 1968, **76**, 291–296.

Lewis, D. J. Partial reinforcement: A selective review of the literature since 1960. *Psychological Bulletin,* 1960, **57**, 1–28.

Lewis, D. C., & Cotton, J. W. The effect of intertrial interval and number of acquisition trials with partial reinforcement on performance. *Journal of Comparative and Physiological Psychology,* 1959, **52**, 598–601.

Lobb, H., & Runcie, D. Intertrial reinforcement as interference with consolidation. *Psychonomic Science,* 1967, **9**, 25–26.

Logan, F. *Incentive.* New Haven: Yale University Press, 1960.

Longstreth, L. E. Partial reinforcement effect and extinction as a function of frustration and interfering responses. *Journal of Experimental Psychology,* 1964, **67**, 581–586.

Ludvigson, H. W., & Sytsma, D. The sweet smell of success: Apparent double alternative in the rat. *Psychonomic Science,* 1967, **9**, 283–284.

Marx, M. H. The need for more specific formulations of the competing response interpretation of extinction. *Psychological Reports,* 1963, **12**, 729–730.

McCain, G. Partial reinforcement effects following a small number of acquisition trials. *Psychonomic Monograph Supplements,* 1966, **1**(Whole No. 12), 251–270.

McCain, G. The partial reinforcement effect after minimal acquisition: Single pellet reward. *Psychonomic Science,* 1968, **13**, 151–152.

McCain, G. The partial reinforcement effect after minimal acquisition: Single pellet reward, spaced trials. *Psychonomic Science,* 1969, **15**, 146–147.

McCain, G. Reward magnitude and instrumental responses: Consistent and partial reward. *Psychonomic Science,* 1970, **19**, 139–141. (a)

McCain, G. The PREE after minimal acquisition: Goal box placement with single pellet reward. *Psychonomic Science,* 1970, **18**, 313–314. (b)

McCain, G., & Brown, E. R. Partial reinforcement with a small number of trials: Two acquisition trials. *Psychonomic Science,* 1967, **7**, 265–266.

McCain, G., & Power, R. Extinction as a function of reinforcement conditions in the start box. *Psychonomic Science,* 1966, **5**, 193–194.

McCoy, D. F., & Marx, M. H. Competing responses and the partial reinforcement effect. *Journal of Experimental Psychology,* 1965, **70**, 352–356.

McHose, J. H. Patterned running as a function of the sequence of trial administration. *Psychonomic Science,* 1967, **9**, 281–282.

McHose, J. H., Jacoby, L. L., & Meyer, P. A. Extinction as a function of number of reinforced trials and squad composition. *Psychonomic Science,* 1967, **9**, 401–402.

Mikulka, P. J., Lehr, R., & Pavlik, W. B. Effect of reinforcement schedules on reward shifts. *Journal of Experimental Psychology,* 1967, **74**, 57–61.

Mikulka, P. J., & Pavlik, W. B. Deprivation level, competing responses, and the PREE. *Psychological Reports,* 1966, **18**, 95–102.

Nelson, P. B., & Wollen, K. A. Effects of ethanol and partial reinforcement upon runway acquisition. *Psychnomic Science,* 1965, **3**, 135–136.

North, A. J., & Morton, M. L. Successive acquisition and extinction of an instrumental response. *Journal of Comparative and Physiological Psychology,* 1962, **55**, 974–977.

Overstreet, L., Jr., & Campbell, P. E. Intertrial reward (ITR) and delay of reward: Effects on resistance to extinction. *Psychonomic Science,* 1969, **17**, 303–334.

Padilla, A. M. A few acquisition trials: Effects of magnitude and percent reward. *Psychonomic Science,* 1967, **9**, 241–242.

Paul, L. High performance obtained with no reward. *Psychonomic Science,* 1969, **14**, 224–225.

Pilley, J. W., & Leeming, F. C. Influence of altered goal-box cues on partial-reinforcement effect. *Psychological Reports,* 1966, **19**, 1259–1266.

Pratt, P. A., & Halpern, J. The effect of extended acquisition upon the replacement capacity of intertrial reinforcement. *Psychonomic Science,* 1969, **14,** 203–204.

Rashotte, M. E., & Surridge, C. T. Partial reinforcement and partial delay of reinforcement effects with 72-hour intertrial intervals and interpolated continuous reinforcement. *Quarterly Journal of Experimental Psychology,* 1969, **21,** 156–161.

Robbins, D. Resistance to extinction as a function of the width of a tube-alley. *Psychonomic Science,* 1966, **6,** 409–410.

Robbins, D. Effect of duration of water reinforcement on running behavior and consummatory activity. *Journal of Comparative and Physiological Psychology,* 1969, **69,** 311–316.

Robbins, D., Chait, H., & Weinstock, S. Effects of nonreinforcement on running behavior during acquisition, extinction, and reacquisition. *Journal of Comparative and Physiological Psychology,* 1968, **66,** 699–706.

Roberts, W. A. Resistance to extinction following partial and continuous reinforcement with varying magnitudes of reward. *Journal of Comparative and Physiological Psychology,* 1969, **67,** 395–400.

Ross, R. R. Positive and negative partial-reinforcement extinction effects carried through continuous reinforcement, changed motivation and changed response. *Journal of Experimental Psychology,* 1964, **68,** 492–502.

Rudy, J. W., Homzie, M. J., Cox, R. D., Graeber, R. C., & Carter, E. N. Effects of sequential manipulations in the within-subjects partial reinforcement runway experiment with rats. *Journal of Comparative and Physiological Psychology,* 1970, **72,** 105–115.

Spear, N. E., Hill, W. F., & O'Sullivan, D. J. Acquisition and extinction after initial trials without reward. *Journal of Experimental Psychology,* 1965, **69,** 25–29.

Spear, N. E., & Spitzner, J. H. Effect of initial nonrewarded trials: Factors responsible for increased resistance to extinction. *Journal of Experimental Psychology,* 1967, **74,** 525–537.

Spence, K. W. *Behavior theory and conditioning.* New Haven: Yale University Press, 1956.

Spence, K. W. *Behavioral theory and learning.* Englewood Cliffs, N. J.: Prentice-Hall, 1960.

Spence, K. W., Platt, J. R., & Matsumoto, R. Intertrial reinforcement and the partial reinforcement effects as a function of number of training trials. *Psychonomic Science,* 1965, **3,** 205–206.

Sperling, S. E. Reversal learning and resistance to extinction: A review of the rat literature. *Psychological Bulletin,* 1965, **63,** 281–297. (a)

Sperling, S. E. Reversal learning and resistance to extinction: A supplementary report. *Psychological Bulletin,* 1965, **64,** 310–312. (b)

Spivey, J. E. Resistance to extinction as a function of number N-R transitions and percentage of reinforcement. *Journal of Experimental Psychology,* 1967, **75,** 43–48.

Spivey, J. E., & Hess, D. T. Effect of partial reinforcement trial sequences on extinction performance. *Psychonomic Science,* 1968, **10,** 375–376.

Spivey, J. E., Hess, D. T., & Aponte, J. F. Modification of reinforcement aftereffects and level of training. *Psychological Reports,* 1968, **22,** 35–42.

Spivey, J. E., Hess, D. T., & Black, D. Influence of partial reinforcement pattern and intertrial reinforcement on extinction performance following abbreviated training. *Psychonomic Science,* 1968, **10,** 377–378.

Spivey, J. E., Hess, D. T., & Black, D. Response decrement as a function of extinction or shift in reinforcement schedule. *Psychonomic Science,* 1969, **13,** 143–144.

Spivey, J. E., Hess, D. T., & Klemic, J. Extinction performance as a function of N-R transitions and intertrial reinforcement with extended partial reinforcement training. *Psychological Reports,* 1968, **22,** 765–771.

Stanley, W. C., & Clayton, F. L. An experimental test of Weinstock's hypothesis concerning the effect of partial reinforcement on extinction. *Psychological Reports,* 1955, **1,** 421–424.

Surridge, C. T. Performance following rewarded and nonrewarded trials at 12-min and 20-sec intertrial intervals. *Psychonomic Science,* 1967, **9,** 501–502.

Surridge, C. T., & Amsel, A. A "patterning" effect that seems unrelated to after-effects from reward and nonreward. *Psychonomic Science,* 1965, **3,** 373–374. (a)

Surridge, C. T., & Amsel, A. Performance under a single alternation schedule of reinforcement at 24-hour intertrial intervals. *Psychonomic Science,* 1965, **3,** 131–132. (b)

Surridge, C. T., & Amsel, A. Acquisition and extinction under single alternation and random partial reinforcement conditions with a 24-hour intertrial interval. *Journal of Experimental Psychology,* 1966, **72,** 361–368.

Surridge, C. T., & Amsel, A. Confinement duration on rewarded and nonrewarded trials and patterning at 24-hour ITI. *Psychonomic Science,* 1968, **10,** 107–108.

Surridge, C. T., Boehnert, J., & Amsel, A. The effect of interpolated extinction on the reacquisition of partially and continuously rewarded responses. *Journal of Experimental Psychology,* 1966, **72,** 564–570.

Surridge, C. T., Rashotte, M. E., & Amsel, A. Resistance to extinction of a running response after a small number of partially rewarded trials. *Psychonomic Science,* 1967, **7,** 31–32.

Sutherland, N. S. Mackintosh, N. J., & Wolfe, J. B. Extinction as a function of the order of partial and consistent reinforcement. *Journal of Experimental Psychology,* 1965, **69,** 56–59.

Tacker, R. S., & Way, J. Motivational properties of nonreward. *Psychonomic Science,* 1968, **10,** 103–104.

Theios, J. The partial reinforcement effect sustained through blocks of continuous reinforcement. *Journal of Experimental Psychology,* 1962, **64,** 1–6.

Theios, J., & McGinnis, R. W. Partial reinforcement before and after continuous reinforcement. *Journal of Experimental Psychology,* 1967, **73,** 479–481.

THEIOS, J., & POLSON, P. Instrumental and goal responses in nonresponsive partial reinforcement. *Journal of Comparative and Physiological Psychology*, 1962, **55**, 987–991.

TRAPOLD, M. A., & DOREN, D. G. Effect of noncontingent partial reinforcement on the resistance to extinction of a running response. *Journal of Experimental Psychology*, 1966, **71**, 429–431.

TRAPOLD, M. A., & HOLDEN, D. Noncontingent partial reinforcement of running: A replication. *Psychonomic Science*, 1966, **5**, 449–450.

WAGNER, A. R. Effects of amount and percentage of reinforcement and number of acquisition trials on conditioning and extinction. *Journal of Experimental Psychology*, 1961, **62**, 234–242.

WAGNER, A. R. Sodium amytal and partially reinforced runway performance. *Journal of Experimental Psychology*, 1963, **65**, 474–477.

WEINSTOCK, S. Resistance to extinction of a running response following partial reinforcement under widely spaced trials. *Journal of Comparative and Physiological Psychology*, 1954, **47**, 318–322.

WEINSTOCK, S. Acquisition and extinction of a partially reinforced running response at a 24-hr. intertrial interval. *Journal of Experimental Psychology*, 1958, **56**, 151–159.

WEINSTOCK, S. A contiguity analysis of learning in the runway. In M. H. Marx (Ed.), *Learning: Theories*. New York: Macmillan, 1970.

WILSON, J. J. Level of training and goal box movements as parameters of partial reinforcement effect. *Journal of Comparative and Physiological Psychology*, 1964, **57**, 211–213.

WILTON, R. Extinction of a runway response following runway or goal box partial reinforcement. *Quarterly Journal of Experimental Psychology*, 1967, **19**, 162–165.

YELEN, R. D. Magnitude of reward and reinforcement aftereffects. *Psychonomic Science*, 1967, **8**, 301–302.

(Received May 11, 1970)

Psychological Bulletin
1971, Vol. 76, No. 6, 432–454

INSTRUMENTALITY THEORIES:
CURRENT USES IN PSYCHOLOGY[1]

TERENCE R. MITCHELL[2] AND ANTHONY BIGLAN

University of Washington

A number of psychologists in a variety of areas have proposed that behavior can be predicted from knowledge of the probability of the behavior leading to various outcomes and the evaluation of these outcomes. Theories based upon this idea are reviewed in the areas of verbal conditioning, attitudes, and industrial psychology. Criticisms and improvements of these theories are discussed.

This study reviews the literature on instrumentality theories in psychology. These theories are distinguished by the hypothesis that the behavior of an individual is in part determined by (*a*) his expectations that the behavior will lead to various outcomes and (*b*) his evaluation of these outcomes. The success of these theories in predicting behavior is evaluated for three distinct areas of research: verbal conditioning, attitudes, and industrial psychology.

The concept of instrumentality is not new to psychology. Both Lewin (1935) and Tolman (1932) emphasized the link between behavior and its perceived outcomes. However, reference to internal, organismic processes as determinants of behavior has not been readily accepted in the United States. This seems due to two closely linked factors. First, stimulus-response theories have dominated psychology in this country, and these theories emphasize factors, external to the organism, that affect behavior. Second, the prevailing behaviorist logic was not adequate for supporting statements containing constructs that refer to internal organismic processes, and there existed no alternate logic that was adequate.

The status of constructs that refer to internal organismic processes in psychology is worthy of a review in itself. We do not discuss the arguments for and against such constructs in detail, but we do present a logic that is felt to be necessary and sufficient for the support of theories that contain constructs referring to internal, organismic processes. Although controversy over the use of such theories has occurred mainly in the learning area (cf. Dulany, 1968; Maltzman, 1966; Spielberger, 1965), the logic we describe has important implications for other areas of psychology.

LOGIC OF THEORETICAL NETWORKS

Dulany (1968) has formulated what he calls a logic of theoretical networks. It draws on the work of Hempel (1952), Carnap (1956), Feigl (1956), and Cronbach and Meehl (1956). At least four ideas may be cited as a basis for this logic. First, Hempel has argued that propositions mentioning unobservables are too weak in isolation to generate verifiable consequences by deduction. This means that the support of propositions about unobservables is possible only when sets of propositions are considered together. Second, confirmation should be substituted for verification as the criterion for the truth of the statements of science. A statement is said to be confirmed when its *deductive implications* are verified; the verification of its deductive implications provides probabilistic support for the statement itself and, thus, confirmation of it. This criterion is considered to be a more accurate construal of the way in which we evaluate the truth of statements. Moreover, it permits the acceptance of statements containing unobservables, since, although they can never be verified, their deductive consequences can.

---

[1] The preparation of this review was supported by Contract NR 177-472, N00014-67-A-0102-0012, Office of Naval Research, Department of the Navy (Fred E. Fiedler, Principal Investigator), and by Contract NR 177-473, N00014-67-A-0102-0013, Advanced Research Projects Agency, Office of Naval Research, Department of the Navy (Fred E. Fiedler, Principal Investigator). The authors are indebted to Gerald R. Oncken and Fred E. Fiedler for their invaluable suggestions.

[2] Requests for reprints should be sent to Terence R. Mitchell, Organizational Research, University of Washington, Seattle, Washington 98105.

Third, a network is said to exist when at least one construct, which is linked to multiple reductions, is linked to another construct, which is also linked to multiple reductions. Each of these links may be formulated in a proposition, and it is this set of propositions that is required for the supportable use of unobservables. Fourth, the adequacy of a theory is evaluated in terms of its competitive support. A theory is competitively supported to the extent that data follow naturally from it, and competing theories require augmenting assumptions or modifications to explain the same data. Hypotheses are also evaluated on the basis of their competitive support.

Dulany's (1968) logic of theoretical networks may be summarized in five principles:

1. *Theoretical terms are introduced postulationally by a network of theoretical sentences.* Each sentence hypothesizes a link between the theoretical terms and another theoretical term or an operation. The hypotheses may originate in common sense, intuition, experimental findings, or any other source the investigator deems useful. Earlier logics treated links between theoretical terms and operations as matters of definition. Following Cronbach and Meehl (1956), the present logic recognizes that we may want to reject an operation as an instance of the construct in question, without changing our idea of the nature of the construct. Hence, links between theoretical terms and operations are hypotheses to be evaluated on the basis of empirical results.

2. *The network is always open to the addition of new terms and to the specification of new linkages between terms that are already in the network.* Thus, the network is never complete.

3. *Terms get their meaning through connections to other variables in the network, which provide the competitive support of the sentences in which they appear.* Thus, the meaning of a term comes from the variables to which it is linked and the supported hypotheses about each of those variables.

4. *The specification of the meaning of any term is always partial.* There are at least two reasons for saying this. First, the linkages of the term to other terms in the network are hypothetical and, thus, are true only with some probability. Second, because the network is open, new terms may always be linked to the term, thereby altering its meaning. This principle implies that we may only converge on the referents and referential meaning of a term. This is in contrast to the behaviorist position, which claims to define terms entirely by the operations to which they are linked (Maltzman, 1966).

5. *Confidence in the referential meaning of a term is as strong as the competitive support of the network or theory in which it appears.* This principle demands that theoretical terms be placed in "richly articulated" networks in order for them to be adequately supported. The greater the number of terms that the term of interest is linked to, and the greater our confidence in this set of linkages, the greater our confidence in the meaning of the term.

What does this logic imply about the use of terms that refer to unobservables? First, terms referring to unobservables may be placed in the network in the same way as any other theoretical terms. Hypotheses concerning the causal status of terms such as expectation or intentions are linkages between the term and other terms. Support for the hypothesis is based on the competitive support of the network in which it appears.

Second, this approach does *not* suggest that constructs referring to unobservables need not be grounded in some form of observable measurement. All terms (both observable and unobservable) are linked, though hypothetically, to operations. Operations are therefore seen as tentative reflections of these constructs. Support for the hypothesis that an operation is a measure of a certain construct is provided to the extent that the measure is related to measures of other constructs in the network in the way that our theory says it should and should not be. Thus, explicit definition of hypothetical unobservables by one or a set of operations is rejected. The bootstrap quality of the development of *both* constructs and their measures is acknowledged in this logic.

Instrumentality theories, as was mentioned before, suggest that behavior is in part caused by (*a*) one's estimate of the probability of the behavior leading to some outcomes and (*b*) the evaluation of these outcomes. Both components of these theories refer to internal processes, and therefore any attempt to support this idea must use a logic that allows the investigator to support both observable and unobservable variables as causes of behavior. We do not mean to imply that all the authors reviewed in the study are proponents of this logic, but rather that this logic is both necessary and sufficient for supporting these theories.

While it is clear that this logic has implications for the controversy concerning the status of cognitive concepts in learning theory, its implications for other areas of psychology may not be as clear. Our review shows that instrumentality theories have been used more extensively in the social and industrial areas of psychology than in learning. This is partly due to the fact that the behaviorist eschewal of cognitive concepts has not prevailed to the same extent in the former areas. However, network logic suggests certain restrictions on the use of cognitive concepts that social and industrial psychology may not always live up to. First, it suggests that cognitive concepts should be linked to multiple operations. This requirement is seldom met. Second, it requires that such terms be placed in "richly articulated" theories. Thus, theories involving only a few constructs are inadequate for the support of statements involving conscious processes.

There are some final restrictions with respect to this review that should be mentioned. First, three distinct antecedents of behavior are conceptualized in the following manner:

$$\underset{\text{environmental events}}{\overset{1}{\text{stimulus or}}} \longrightarrow \underset{\text{and weighting}}{\overset{2}{\text{cognitions, evaluations,}}} \longrightarrow \overset{3}{\text{intentions}} \longrightarrow \overset{4}{\text{behavior}}$$

This article reviews studies that use instrumentality as their theoretical base for explaining relationships between 2 and 3, as well as between 2 combined with 3, and 4. Therefore, studies that examine just the relationship between 3 and 4 (e.g., Locke, Bryan, & Kendall, 1968; Triandis, 1970) are not included.

A second area of related work that is not reviewed is that of behavior decision theory. These theories predict that when an individual chooses among various alternatives, he chooses the alternative that he perceives will maximize his gain. Perhaps the most well-known theory in this area is the subjective expected utility (SEU) theory (Edwards, 1961, 1962), which predicts that a given choice of alternatives can be predicted from the knowledge of the subjective probabilities (SPs) that certain outcomes will follow the act (SP_i) multiplied by the subjective evaluation (utilities, U) of the outcomes (U_i). The products are summed over all the outcomes to give the SEU for a given choice: $SEU = \Sigma SP_iU_i$, and the theory predicts that the alternative for which SEU is greatest will be chosen.

Review of this work is omitted for a number of reasons. First, much of this material has been reviewed elsewhere (Becker & McClintock, 1967; Edwards, 1961, 1962; Peterson & Beach, 1967). Second, a large amount of this work is concerned with the proper use and measurement of subjective probabilities (Beach & Wise, 1969a, 1969b; Peterson & Miller, 1965). While it is recognized that the measurement issues are important for both theoretical and empirical reasons, they are not central to evaluating instrumentality theories' ability to account for behavior.

VERBAL CONDITIONING

Theory of Propositional Control

Dulany (1961, 1962, 1968) has developed a theory that includes instrumentalities. The theory is intended to account for behavior in verbal conditioning and concept acquisition experiments. The theory specifies the propositions that precede the subject's response. It is concerned with the immediate antecedents of specific behavioral responses and not with the prediction of behavior in general. Dulany (1968) expressed the central equation of his theory as follows:

$$B \approx BI = [(RHd)(RSv)]w_0 + [(BH)(MC)]w_1,$$

where

B = behavior;
BI = the subject's intention to make a particular response or class of responses;

RHd = a "hypothesis of the distribution of reinforcement"; that is, the subject's hypothesis that the occurrence of the particular response will lead to a certain event or class of events;
RSv = the subjective value of the reinforcement; that is, the subject's evaluations of (or attitudes toward) those events;
BH = The subject's "behavioral hypothesis"; that is, his belief as to what he is expected to do, or what he should do in the situation;
MC = the subject's "motivation to comply"; that is, how much the subject *wants* to do what he believes is expected of him;

w_0 and w_1 = beta weights, which may take values between -1.00 and $+1.00$.

This theory says that a subject's behavioral intention is determined by an instrumentality component and a social component. The instrumentality component is (RHd)(RSv). RHd is the subject's hypothesis about the distribution of reinforcement; that is, his perception of the relationship between his response and the consequences of the response. It is multiplied by RSv, the subject's evaluation of the consequences. The second component deals with the normative or social influences on the subject's behavioral intention: (BH)(MC). BH is the subject's behavioral hypothesis, what he thinks he is expected to do in the situation. This term is multiplied by MC, which is the subject's motivation to comply with what he thinks he is expected to do.

The subject's actual response is caused in part by his behavioral intention (BI). As Dulany uses the term, BI refers to a specific intention to engage in a specific act in a given situation. Other investigators have used this term to refer to general response sets, which may be evoked in a variety of situations (Triandis, 1964).

According to Dulany's theory, situational, personality, and attitude variables affect behavior and behavioral intention only through their effect on the social and normative components of the theory. For example, reinforcement contingencies affect the subject's hypothesis about the distribution of reinforcement (RHd) and his behavioral hypothesis (BH). These in turn affect behavioral intention in the manner specified above.

In practice, Dulany used multiple linear regression to estimate the weights for the two components. This means that the contributions of the two components to the prediction of behavioral intention and behavior can vary over situations, behaviors, and individuals. The weights are mathematically determined, depending upon the best empirical fit to predict behavior. For example, in situations of high group salience, the normative component may account for most of the variance in behavioral intention and behavior.

Empirical Support

Two studies (Dulany;[3] Dulany, Schwartz, & Walker[4]) provide support for this theory. In the first study, Dulany related experimental manipulations and the components of his theory to behavior in a verbal conditioning experiment. Female subjects were seated in a 12 × 7 foot chamber that was kept at a constant temperature of 110 degrees centigrade (C.) and a relative humidity of 35%. The subject was told that she would perform a verbal task under simulated climatic conditions. The subject's head was kept stationary with a headrest so that a stream of air from a nearby tube could always arrive at the side of her face. The task consisted of choosing and reading one of two sentences printed on a card, for 100 cards.

For example, one card contained the following:

> The frog suddenly hopped along a slick log and disappeared into a pond.
> The frog quickly hopped over the slippery log and vanished into a pond.

Another example is:

> The frog quickly hopped along the slick log and disappeared into the pool.
> The frog suddenly hopped over a slippery log and vanished into the pond

---

[3] Dulany, D. E. The separable effects of the information and effect conveyed by a reinforcer. Paper presented at the meeting of Psychonomic Society, St. Louis, October 1964.

[4] Dulany, D. E., Schwartz, S., and Walker, C. Why the affective and distributional parameters of reinforcement interact. Paper presented at the meeting of the Psychonomic Society, Chicago, October 1965.

In all cases, a stream of air followed the reading of sentences containing the articles "a" before "slippery" or "slick." Since each of the other words in these sentences appeared with the word "a" with a certain frequency, the probability of obtaining a blast of air for using each word could be specified. Table 1 presents the probability of obtaining a blast of air that was associated with the use of each of the words when taken over all sentences.

The design was a 3 × 3 factorial corresponding to three levels of reinforcement value and three conditions of instruction about the significance of the reinforcement. There were 12 subjects in each of the resulting nine cells. The air blast associated with the use of "a" could be 40 degrees cooler than chamber temperature, at chamber temperature, or 40 degrees warmer than chamber temperature. That these reinforcements differed in value was substantiated by ratings of pilot subjects. On an 11-point bipolar scale that ranged from + 1 (pleasant) to − 1 (unpleasant), mean values were .81, − .02, and −.84 for the three conditions of reinforcement. The three instruction conditions consisted of telling subjects that the blast of air (*a*) meant that they used the sentence correctly, (*b*) had nothing to do with the sentence they chose, or (*c*) meant that they had chosen the wrong sentence.

The behavior to be predicted was the subject's responses on the last 20 of 100 trials. According to Dulany's theory, the experimental conditions of the experiment should influence behavior only through their influence on the two components of the theory and the influence of these components on behavioral intention. This means that experimental conditions should not be related to final performance when the propositional control variables are partialed out.

TABLE 1

PROBABILITY OF RECEIVING BLAST OF AIR ASSOCIATED WITH WORDS IN THE DULANY EXPERIMENT

Attributes	Probability value
a	+1.0
suddenly	.9
along	.8
slippery	.7
pool	.6
disappeared	.5
vanished	.5
pond	.4
slick	.3
over	.2
quickly	.1
the	0.0

After completion of 100 trials, subjects were led from the room and given a questionnaire to assess their awareness over the last 20 trials. Subjects' behavioral intention as well as the other parts of the equation were measured with a questionnaire. Thus, subjects' hypotheses about the distribution of the reinforcer (RHd) and their evaluation of the reinforcer (RSv) were assessed for the first component of the equation, and what subjects thought they should do in the situation (BH), and their desire to comply with what they thought they should do (MC) were measured. Care was taken to present questionnaire completion as an essential part of the experiment.

Two experimenters were used; subjects were tested from early winter to late spring, and the positions of questions in the questionnaire were scrambled. None of these factors influenced the variables in the equation.

The results support Dulany's theory. The multiple correlation between behavioral intention and the two products (RHd × RSv) and (BH × MC) was .88. The beta weights for the two predictors were .19 and .81, respectively, indicating that in this situation, subjects' hypotheses about what they are supposed to do are the more important determinant of their behavioral intentions. Obviously, for this approach to have any value, behavioral intention must be correlated with behavior. The correlation in this case was .94.

The results of this study underscore the importance of cognitive variables in accounting for the behavior of human subjects. Taken together, the two experimental manipulations accounted for 36% of the variance in actual behavior. This is in contrast to the 88% of response variance, which was accounted for by behavioral intention.

We suggested above that if Dulany is correct that experimental manipulations influence behavior through their effect on cognitions, then the relationship between manipulations and behavior should be zero when the relevant cognitive terms are partialed out. This was, in fact, the case; experimental manipulations were not significantly related to behavior when BI and RHd were "at zero or partialed out."

Although these results support Dulany's theory, they do not strongly support the instrumentality hypothesis, namely, that a person's behavior is the result of its perceived consequences and the evaluation of those consequences. A multiple correlation between the components of the theory and behavior was not reported. However, in predicting behavioral intention, the beta weight for the instrumentality component (RHd)(RSv) was only .19, and the correlation coefficient was .40. These data indicate that the perceived instrumental relations in this situation were not substantially related to behavioral intention. The beta weight of .81 for the normative component (BH)(MC) showed that the subjects' perceived expectations were a more potent influence on his intentions than were the instrumentalities.

Dulany et al. (see Footnote 4) conducted a study similar to the one just described in order to see if the instrumentality component was more highly related to behavioral intention and behavior when subjects' behavioral hypotheses were fixed at zero. Fixing BH at zero fixes the (BH)(MC) component at zero and brings the contribution of this component to the prediction of BI to zero. Relevant experimental procedures were the same as those described above, except that all subjects were instructed that the air blast had nothing to do with the sentence that they chose to read. For subjects whose behavioral hypothesis was at zero, the correlation between (RHd)(RSv) and BI was .72. The correlation between BI and behavior was .90 in this study.

Discussion

Taken together, these studies provide substantial support for Dulany's theory, but only partial support for the basic instrumentality conception. It appears that in many situations, normative or social factors are more important for the subjects' intentions and behavior than are the subjects' perceived instrumentalities.

Dulany's use of multiple regression presents two problems. First, its use in the prediction of BI and behavior from the theory of propositional control demands that all subjects weigh the two factors (RHd)(RSv) and (BH)(MC) in the same way. That is, the empirically determined weights are the mathematical best fit for a set of subjects. Thus, prediction of behavioral intention and behavior will be attenuated to the extent that different people weigh these factors differently. A second problem with the use of multiple regression is that the weights have no psychological meaning. The theory could be improved by developing a weighting system that reflects the subjects' own method of weighting components.

Finally, the attempt to generalize this theory to less specific situations may prove to be difficult. That is, one of the major problems with the above approach is that the components of the equation change over time. In the studies mentioned above, the experimenter is asking subjects to recall information about what they thought over the last 20 of 100 trials. In more general situations where one is attempting to predict, for example, work behavior, one may find that stable estimates of the variables of Dulany's theory are hard to derive.

ATTITUDES

Instrumentality theories have been useful in the attitude area in two ways. First, they have suggested a way of conceptualizing attitudes. Second, they have contributed to the prediction of behavior from attitudes.

Peak (1955) first pointed out that attitudes toward any object are dependent upon the object's utility for attaining goals. More specifically, she stated that attitudes are related to (*a*) their instrumentality for aiding or interfering with goal attainment and (*b*) the satisfaction derived from reaching those goals. Instrumentality is defined as the perceived relationship between the attitude object and the attainment of goals.

A test of this attitude theory was provided by Rosenberg (1956). He hypothesized that attitude is related to the algebraic sum of the products obtained by multiplying (*a*) the importance of each of a set of values by (*b*) the instrumentality of the attitude object for achieving or blocking realization of each value.

Symbolically, this hypothesis may be stated:

$$A = \sum_{1}^{n} I_i V_i,$$

where

A = attitude;
I_i = the rated instrumentality of the attitude object for attaining or blocking each of a set of values;

and

V_i = the rated importance of the value;
n = number of values.

Subjects' attitudes toward the issue of allowing members of the Communist Party to address the public were assessed by means of a standard attitude questionnaire. This questionnaire was of the Likert type (Likert, 1932). Test-retest reliability over a 2-month period was reported to be .72. Three to 5 weeks after completion of the Likert attitude scale, subjects rated the instrumentality of allowing Communists to speak for attaining each of 35 personal values and the importance to them of these values as a source of satisfaction. Each value could be rated from +10 to −10 on the importance dimension. The instrumental relationship between letting Communists speak and each value could be rated from + 5 to − 5. The sum of the products over the 35 values was significantly related to the original measure of attitude, $\chi^2 = 26.33$ ($p < .001$). A second score was computed in the same manner for values that were rated by the subjects to be most salient. This score was even more strongly related to the original attitude measure ($\chi^2 = 45.09$; $p < .001$). This result suggests that attitudes may depend on only a small number of values that are important to the individual. Using a standard set of questions for all subjects may be adding superfluous or even confusing information.

Rosenberg (1956) reported successful replication of these results for the attitude "allowing Negroes to move into white neighborhoods." He did not report the details of the study, except to say that it was based on "identical measurement operations" and a parallel analysis.

Fishbein (1963, 1965b, 1967a, 1967b) has also argued that instrumentalities are related to attitudes. The formula for his attitude theory is expressed as follows:

$$A_o = \sum_{i=1}^{n} B_i a_i,$$

where

A_o = the attitude toward the object o;
B_i = the strength of belief i about o; that is, the "probability" or "improbability" that o is associated with some other concept x_i ;
a_i = evaluative aspect of B_i; that is, the evaluation of the related concept x_i;
n = the number of beliefs about o; that is, the number of responses in the individual's habit-family hierarchy.

Fishbein based his theory partly on an analysis of existing attitude measurement techniques. According to him (1965b), the most frequently employed techniques (Thurstone Scale, Likert Scale, Semantic Differential) measure subjects' beliefs about the association between the attitude object and other objects, events, or values; and their evaluation of these objects, events, or values. For example, on the semantic differential, subjects are asked to rate the attitude object on a series of 7-point (− 3 to + 3) bipolar adjective scales. The adjectives that are typically used were determined by Osgood, Suci, and Tannenbaum (1957) to be heavily loaded on an evaluative dimension, so that one adjective reflects positive evaluation of the attitude object and the other negative evaluation (e.g., good–bad). According to Fishbein, the subject's check on the 7-point scale indicates whether he thinks the attitude object is associated with a positive or negative state (a_i) and his belief about the strength of association (B_i). Thus, when the subject checks + 2 on the good–bad scale, he is indicating that the attitude object is associated with a positive state and that the strength of the association is 2. The subject's total score on this instrument is conceptualized as the sum of (*a*) his beliefs about the relationship between the attitude object and certain positive or negative states ($a_i = +1$ or -1) multiplied by (*b*) his perception of the strength of these relationships ($B_i = 0$ to 3) (M. A. Fishbein, personal communication, June 1969).

To provide support for his theory, Fishbein (1965b) has compared subject scores on semantic differential scales with scores that were directly based on the subjects' perceptions of the instrumental relations between the attitude object and other objects. He had 125 subjects list what they believed to be five characteristics of Negroes. The 10 most frequently cited characteristics were selected, and belief statements were constructed for each one (e.g., Negroes have dark skin). Two weeks later, 50 subjects rated each characteristic on five bipolar evaluative adjective scales (e.g., good–bad), and each belief statement on five probabilistic bipolar scales (e.g., probable–unprobable). Each subject also rated the concept "Negro" on the five bipolar evaluative scales. The correlation between $\sum_{i=1}^{n} B_i a_i$ and the direct evaluation of "Negro" was .801 ($p < .001$). Fishbein reported similar support for leaders' attitudes about group members (Fishbein, 1965a), and voter's attitudes about political candidates (Fishbein & Feldman).[5] Thus, both Rosenberg's and Fishbein's research provide good evidence that attitudes depend on the perceived instrumental relations between the attitude object and other objects or values and the evaluation of these related objects or values.

Attitudes and Behaviors

Fishbein (1967b) also has developed a theory of the attitudinal and normative antecedents of behavioral intention and behavior. Attempts to predict specific behavior from attitudes have typically led to nonsupportive results (LaPiere, 1934, is perhaps the most famous example). Fishbein (1967b) pointed out that there are at least two reasons why this should be so. First, we have often measured attitude toward a general stimulus object or class of objects rather than a specific object. Second, the behavior predicted may be only partially related to attitude. That is, people may have the same amount of affect toward an object, and yet behave differently toward it (Doob, 1947). Because of these considerations, Fishbein's theory assesses specific attitudes about specific behaviors, and also considers normative influences on behavior.

Fishbein's theory is essentially an adaptation of Dulany's theory of propositional control. The basic equation is:

$$B \approx BI = [\sum_{i=1}^{n} B_i a_i]w_0 + [(NB)(MC)]w_1,$$

where

B = behavior;

BI = intention to make a particular response;

B_i = beliefs about the consequences of performing a particular behavior (in a given situation); that is, the probability or improbability that the performance of behavior X will lead to some consequence y_i;

a_i = the evaluative aspect of B_i; that is, the subject's evaluation of y_i, the consequences;

NB = a normative belief; that is, the subject's expectation about what he *should* do in this situation;

MC = his desire or lack of desire to do what he thinks he should do; that is, his motivation to comply with the norm;

n = number of beliefs;

w_0, w_1 = empirically determined weights.

However, as stated earlier, Fishbein considered that

$$A_o = \sum_{1}^{n} B_i a_i,$$

where A_o = the attitude toward the subject o, and since, in this case, o is a specific act, he considers the following to be an equivalent formulation of his theory of behavior;

$$B \approx BI = [A_{act}]w_0 + [(NB)(MC)]w_1,$$

where A_{act} = attitude toward the act or specific behavior. Thus, the first component (A_{act}) is conceptualized as the instrumentality part of the equation. In the three studies reviewed in the following section, three different techniques were used to measure this component:

---

[5] Fishbein, M., and Feldman, S. Social psychological studies in voting behavior: I. Theoretical and methodological considerations. Paper presented at the annual meeting of the American Psychological Association September 1963.

1. One scale assessing whether the act leads to "good or bad consequences," and one scale for the subjects' confidence or certainty of this relationship.
2. Four bipolar adjective scales as an estimate of $\sum_{1}^{n} B_i a_i$.
3. Scales assessing the subjective probabilities of two choices leading to each of the four outcomes and the utility of each outcome in a prisoner's dilemma game.

To justify the interchangeability of these different measures, it should be shown that they are highly correlated and that they are related to other variables in the theory in the same way. Support for the convergence of the bipolar scales and measures of subjects' perceived instrumentalities was reviewed above. In one of the studies reviewed in the next section (Ajzen & Fishbein, 1969), the authors reported that two different measures of the instrumentality component (Measures 2 and 3) were related to other variables of Fishbein's theory in the same manner. However, they only present the relevant relationships between the bipolar adjective measure and the other variables of the theory.

The second component of Fishbein's theory (NB)(MC) is essentially the same as that used by Dulany. Fishbein's NB and Dulany's BH both refer to the subject's belief about what he should do in a given situation. MC stands for the subject's motivation to comply in both these theories.

Empirical Support

Fishbein (personal communication, June 1969) first used this theory to predict the occurrence of premarital intercourse for a group of 14 female and 21 male students. In this study, the instrumentality component (A_{act}) of the theory was measured with two questions. The first asked subjects whether they felt that sexual intercourse led to good or bad consequences; the second asked them to rate their certainty about the relationship between sexual intercourse and its consequences. The second component of Fishbein's theory was measured by asking subjects to indicate what they believed they should do (NB) and their motivation to comply with what they thought they should do (MC). Finally, subjects' behavioral intentions were assessed. At the end of the semester, subjects were asked whether they had engaged in premarital sexual intercourse. To preserve anonymity, subjects identified themselves only by indicating their mother's maiden name.

Using the two components of the theory, Fishbein computed the multiple correlations for predicting actual behavior and behavioral intentions. These results are presented in Table 2. They indicate that the theory predicts females' intentions and behavior very well. The behavior of males was not significantly related to the components of Fishbein's theoretical equation, although the intentions were. The instrumentality component played a significant role in the subjects' behavioral intention; A_{act} correlated .52 ($p < .05$) with males' BI, and .92 ($p < .05$) with females' BI. Also, A_{act} correlated .72 ($p < .05$) with the females' behavior, and .26 (*ns*) with the males' behavior. Since the number of subjects was small, and the behavior to be predicted was a dichotomous variable, these results constitute only fair support for the theory.

One problem pointed out by this study was that dyadic situations have different parameters than situations in which the subject is by himself. When the behavior in question (premarital sexual intercourse, in this case) is contingent upon the actions of another, intentions may be only partially related to actual behavior.

A second study by Ajzen[6] used Fishbein's theory to predict communication and compliance in 48 three-person groups. Each group member was seated at the corner of a triangular board whose corners could be raised or lowered by the subject. Information about the position of the board was provided by spirit levels that were placed at each corner. The task consisted of balancing the board. Partitions prevented subjects from seeing each other, but they were allowed to communicate with written messages.

---

[6] Ajzen, I. Attitudes, normative beliefs, and the prediction of behavior: An empirical investigation. Unpublished master's thesis, University of Illinois, Urbana, 1967.

TABLE 2

PREDICTION OF PREMARITAL SEXUAL INTERCOURSE

Sex	Multiple R using both components		Correlation with A_{act}	
	BI	B	BI	B
Females	.93*	.83*	.92*	.72*
Males	.86*	.29	.52*	.26

Note.—B = behavior, BI = behavioral intention.
$^{*}p < .05$.

Each subject could raise or lower his corner one step, and communicate with one other subject in a given trial (i.e., instruct him to raise or lower his corner). The behavior to be predicted was the number of instructions sent by each subject to Member 1 (on his left) and Member 2 (on his right), as well as the number of times he complied with messages sent by the other two members for 20 experimental trials.

Subjects completed 10 trials in order to acquaint themselves with the procedure. A pretest questionnaire was then administered. Twenty more trials were then completed, and a final posttest questionnaire distributed. The questions on the pretest asked subjects about their intended behavior on trials that were to follow. The instruction for the posttest required the subjects to answer the question as if they were going to participate in the experiment again. Both the pretest and posttest gathered the following information:

1. A_{act}. Attitudes toward the two acts to be predicted. These attitudes were measured on four bipolar adjective scales.
2. NB. Perceived expectations of the other two members with regard to the two acts.
3. MC. Motivation to comply with the expectations of the other two members.
4. BI. Behavioral intentions with regard to the two acts.

The multiple correlation coefficients for predicting BI from A_{act} and NB were generated both for the pre- and posttest questionnaires to predict both the number of instructions sent and the number of times the subject complied with his two other partners. For these multiple correlations, $N = 144$. Note that A_{act} was not measured by obtaining the probability of the act leading to a set of outcomes and the evaluation of these actions. We discussed in the previous section the problems involved in substituting a standard measure for more direct measures of subjects' perceived instrumentalities. Also note that motivation to comply (MC) was omitted from these analyses. The author reported this variable did not increase, and often attenuated the multiple correlation coefficients and was therefore dropped from the analysis. The coefficients for predicting BI from A_{act} and NB are presented in Table 3.

TABLE 3

MULTIPLE CORRELATION COEFFICIENTS FOR PREDICTING BI FROM A_{act} AND NB IN BALANCE BOARD STUDY ($N = 144$)

Measure	Prequestionnaire		Postquestionnaire	
	Member 1	Member 2	Member 1	Member 2
Instructions	.692*	.720*	.796*	.910*
Compliance	.647*	.702*	.898*	.839*

Note.—BI = behavioral intention, A_{act} = attitude toward the act or specific behavior, NB = normative belief.
$^{*}p < .01$.

These results show that there are substantial relationships between some of the variables of the theory and subjects' behavioral intention. However, since MC was dropped from the analysis, the results cannot be considered as good support for Fishbein's theory. Indeed, to the extent that the inclusion of MC attenuated the multiple correlations, the theory is disconfirmed. Unfortunately, Ajzen (see Footnote 6) did not report the multiple correlations that were obtained when MC was included.

TABLE 4

CORRELATIONS OF BI WITH A_{act} IN BALANCE BOARD STUDY ($N = 144$)

Measure	Prequestionnaire		Postquestionnaire	
	Member 1	Member 2	Member 1	Member 2
Instructions	.409*	.568*	.524*	.802*
Compliance	.605*	.652*	.812*	.787*

Note.—BI = behavioral intention, A_{act} = attitude toward the act or specific behavior.
$^{*}p < .01$.

Correlations between A_{act} and BI are presented in Table 4. These results support the

TABLE 5

CORRELATIONS BETWEEN SPECIFIC BI AND B IN BALANCE BOARD STUDY ($N = 144$)

Measure	Prequestionnaire		Postquestionnaire	
	Member 1	Member 2	Member 1	Member 2
Instructions	.690*	.690*	.822*	.883*
Compliance	.126*	.246*	.446*	.526*

Note.—B = behavior, BI = behavioral intention.
* $p < .01$.

contention that attitudes toward specific behaviors are related to behavioral intention. And, if Fishbein is correct that attitudes involve instrumental relations, these results may be taken as support for the instrumentality hypothesis. Correlations between A_{act} and behavior were not reported.

The correlations between specific behavioral intentions (BI) and actual behavior are presented in Table 5. The variance accounted for in compliance behavior is low for both the pre- and postquestionnaires (0% to 25%). The authors suggested that one problem with the compliance measure was that some subjects received very few or no opportunities to comply. A subject could not comply, even if he wanted to, unless he received an instruction. Therefore, a second analysis examined the compliance behavior just toward the member who sent the greatest number of instructions. The correlations between BI and B for the pre- and postquestionnaires were increased to .242 and .582 ($N = 144$), respectively. Even though both correlations are significant ($p < .01$), these modifications still leave large amounts of variance unaccounted for.

The difference between pre- and postquestionnaire responses suggests an additional problem for the theories reviewed in this study. It is clear from the data that the posttest responses did a better job of predicting BI and B than did the pretest responses. A number of possible reasons for this discrepancy were presented by the author. First, an individual's BI may change during the course of the experiment. In other words, the behavior of others may change what he intends to do. Second, the subject's competence on the task may change his intention, and, third, the subject may attempt to report his intentions on the posttest as more in line with what he actually did—a "cognitive bolstering" process. If cognitive bolstering is occurring, then it is unlikely that the cognitions that are measured in this theory could be *causing* behavior. In any case, it appears that some of the components of the theory are not stable over time. To the degree that this is the case, the applicability of the theory is, of course, limited.

More recently, Ajzen and Fishbein (1969) examined the ability of the theory to predict behavior in prisoner's dilemma games. In the two games used by Ajzen and Fishbein, each of the two players accrue points by choosing between two alternative responses. For each game, a matrix indicated the number of points on each trial which each player could obtain, depending on his response and his partner's response. More specifically, each player chooses C or D and the combination of these choices determines a cell in the matrix:

$$\begin{array}{cc|cc|cc|} & & \multicolumn{4}{c}{\text{Player 2}} \\ & & \multicolumn{2}{c}{C_2} & \multicolumn{2}{c}{D_2} \\ \hline \text{Player 1} & C_1 & +9, & +9 & -10, & +10 \\ \hline & D_1 & +10, & -10 & -9, & -9 \\ \hline \end{array}$$

Game 1

$$\begin{array}{c|cc|cc|} & \multicolumn{2}{c}{C_2} & \multicolumn{2}{c}{D_2} \\ \hline C_1 & +1, & +1 & -12, & +12 \\ \hline D_1 & +12, & -12 & -1, & -1 \\ \hline \end{array}$$

Game 2

The first number in each cell indicates the points which the first player receives, and the second number indicates the points for the second player. For example, in Game 1, if Player 1 chooses Response C, and Player 2 chooses Response D, then Player 1 loses 10 points, and Player 2 gains 10 points. Alternative C is defined as a cooperative response, and D as a competitive one. If a player chooses C in Game 1, he stands to gain + 9 or lose − 10, depending on his partner's response. If he chooses D, he stands to gain + 10 or lose − 9.

However, only if both subjects choose C can they both win points on a given trial (cooperation).

The matrices for the two games differ in the way that payoffs are distributed; the matrix for the first game provides far less pressure toward competition between players. For a more detailed description of these games, see Rapoport and Chammah (1965).

Subjects were 96 undergraduates. Each subject played both games. The first game was played for eight trials and was followed by the administration of a questionnaire. After completing the questionnaire, 10 more trials of the first game were played. The same process was repeated for the second game. The number of cooperative moves on the last 10 trials of each game was the measure of behavior.

The questionnaire assessed each of the following:

1. A_{act}, the attitude toward cooperation. This was measured on four bipolar adjective scales.

2. Subjective conditional probabilities (SP), the subject's perception of the likelihood of his partner choosing each alternative given each of the choices he could make, for example, the probability of his partner choosing D, given that he chose C.

3. Payoff utilities (U). An evaluative measure of each payoff.

4. Normative beliefs (NB). A measure of what is perceived to be expected by one's partner.

5. Motivation to comply (MC). A measure of the subject's willingness to comply with the perceived expectations of his partner.

It will be recalled that the theoretical equation may be written:

$$B \approx BI = [A_{act}]w_0 + [(NB)(MC)]w_1$$

and that Fishbein considers A_{act} to be the instrumentality component. This test of the theory employs two measures of the first component: an estimate of A_{act} in the form of four bipolar scales and a measure of the subjective expected utility (SEU) of cooperation. SEU was computed by summing the products of the subjective probability (SP) of each act times its utility (U): $SEU = \Sigma SP_i U_i$. The measure of A_{act} and SEU correlated .63 for the first game and .67 for the second. These relationships do not appear to be large enough to justify the conclusion that both measures reflect the same construct. We indicated above, however, that these variables had similar relationships to other variables of the theory.

As in the study of the triangle board, MC was dropped from the analysis because it lowered the predictive ability of the equation. Thus, NB was the only variable included in the normative component of the theory.

The authors reported correlations between BI and B of .85 and .84 for Games 1 and 2, respectively. The multiple correlations for predicting BI from A_{act} and NB were .73 and .79 for the two games. The correlations between A_{act} and B were .63 and .70. Table 6 presents these results.

These results are consistent with those reported in earlier studies. However, since the motivation to comply variable was dropped from the analysis, these results do not constitute support for Fishbein's original theory. It appears that the normative component of the theory needs revision. Second, the authors reported SEU to be more highly related to actual behavior than A_{act} in both Games 1 and 2 (.70 to .63 in Game 1, and .75 to .70 in Game 2), and yet the A_{act} estimate of the first component is used throughout their report of the study. If attitude toward an act is indeed determined by the probability of the act leading to certain outcomes and an evaluation of those outcomes, why not use this variable instead of an estimate? The authors argued that A_{act} should be a better estimate because it takes into account a broader number of possible consequences of the given act. How-

TABLE 6

PREDICTION OF COOPERATIVE RESPONSES IN THE PRISONER'S DILEMMA GAME

Correlation	Game 1	Game 2
Correlation of A_{act} and BI	.75*	.74*
Multiple *R* for BI	.89*	.85*
Correlation of A_{act} and B	.63*	.70*
Correlation of SEU and B	.70*	.75*
Multiple *R* for B	.73*	.79*

Note.—A_{act} = attitude toward the act or specific behavior, BI = behavioral intention, B = behavior, SEU = subjective expected utility.
* $p < .01$.

ever, in the above study, SEU was at least as strongly related to B as was A_{act}. The correlation between BI and SEU is not reported. These discrepancies require clarification.

Discussion

Instrumentality theories have been used in attitude research in two ways: First, attitude itself has been defined as the probability of the attitude object leading to various outcomes or values and the evaluation of these outcomes. Both Fishbein (1965b) and Rosenberg (1956) presented good support for this hypothesis. Second, Fishbein has included the above theory of attitude in a more complex theory designed to predict social behavior. This latter theory, which includes both attitudinal (instrumentality) and normative predictors of behavior, has received moderate support.

One point brought out by tests of this theory is that social behavior is often dependent upon other people's behavior. What one does in social situations is contingent upon what others are doing or might do. Only in the last study reviewed (Ajzen & Fishbein, 1969) was an attempt made to measure this interdependence (i.e., the subjective conditional probabilities), and this measure was then dropped from the analysis.

The normative component also presented problems. The motivation to comply variable did not interact with normative beliefs, as the theory predicted that it should. The fault may be with the conceptualization or the measurement of NB and MC, or both.

In Tables 2, 3, 4, and 6, the multiple correlations using both components to predict BI or B can be compared to the zero-order correlations between the instrumentality component and BI or B. All the correlations between BI or B and the instrumentality component are significant except for the correlation with males' behavior in the study predicting premarital sexual intercourse. Also, the differences between these multiple correlations and zero-order coefficients are relatively small, indicating that in these situations, the instrumentality component controls a substantial amount of the predictable variance.

INDUSTRIAL PSYCHOLOGY

Investigators in industrial psychology have been less reluctant to employ constructs that refer to internal, organismic processes. Accordingly, instrumentality theory has been used extensively in this area.

Georgopoulos, Mahoney, and Jones (1957) first tested an instrumentality hypothesis in the industrial area. They suggested that worker productivity could be predicted from the worker's perception of the degree to which productivity is a path to the attainment of personal goals. If the worker sees high productivity as a path to his goals, then he will be a high producer; if he sees low productivity as a path to these goals, then he will be a low producer.

Georgopoulos et al. tested this hypothesis in a household appliances company. They obtained measures of the perceived instrumentality of high and low productivity for each of 10 goals. Their sample of 621 workers was split into high- and low-need groups on the basis of their ranking of three of these goal items: "making money in the long run"; "getting along well with the work group"; and "promotion to a higher base rate." Thus, the high-need group ranked these items as higher in importance to them than did the low-need group. The measure of productivity was based on subject reports of the percentage above or below the company-set standard they had reached. This measure was the form in which productivity was typically dealt with in the plant; hence, it was not alien to the workers. Finally, workers were split into high- and low-freedom groups on the basis of their ability to set their own work pace (if they could, they were in the high-freedom group), their age (20–59 years was high-freedom group), and their amount of experience (more than 6 months was high-freedom group).

The results of this study support the basic goal path or instrumentality hypothesis. First, high productivity was associated with the perception that high productivity leads to each of the three goals in question. Second, for workers who ranked a goal as being important to them personally, the relationship between productivity and perceived instrumentality was stronger. Third, the relationship between perceived instrumentality and productivity was significantly higher for those who were classified as "free." Finally, for workers who were both "free" and had a high need on an item, there was a stronger relationship between perceived instrumentality and productivity than for workers with any other combination of these two factors. To summarize, the above study supports the hypothesis that productivity is related to its perceived instrumentality for the attainment of goals and that degree of attraction to the goals and freedom of action mediate this relationship.

In a similar study, Porter and Lawler (1968) examined attitudes toward pay as they related to the performance of managers in industrial and governmental organizations. They compared the performance of the third of their sample that perceived pay as a probable outcome of performance with the third that saw little relation between performance and pay. Performance (as rated by the subjects and their superiors) was significantly higher for the former group. When samples were split according to the rated importance of pay to the subjects, the relationship between performance and the expectancy that pay results from performance was stronger for the high-importance group than for the low-importance group. This suggests that performance is highest for those who perceive it as leading to *valued* outcomes and is consistent with the Georgopoulos et al. (1957) study.

A more complete statement of instrumentality theory in industrial psychology was presented by Vroom. His statement of the theory has been the basis for most of the subsequent work in this area. Vroom (1964) presented two models, the first for the prediction of the valences of outcomes, and the second for the prediction of force toward behavior. An outcome is simply anything an individual might want to attain. The valence of an outcome for a person is defined conceptually as the strength of his positive or negative affective orientation toward it. Similar to Lewin's use of the term, valence refers to the anticipated satisfaction associated with an outcome, and is distinguished from the value of the outcome—the actual satisfaction resulting from attainment of the outcome.

The valence model states that the valence of an outcome to a person is a monotonically increasing function of the algebraic sum of the products of the valences of all other outcomes and his conceptions of the specific outcome's instrumentality for the attainment of these other outcomes. Symbolically,

$$V_j = f \sum_{k=1}^{n} (V_k I_{jk}),$$

where

V_j = the valence of outcome j;
I_{jk} = the cognized instrumentality of outcome j for the attainment of outcome k;
V_k = valence of outcome k;
n = the number of outcomes.

Cognized or perceived instrumentality is defined conceptually by Vroom (1964) as the degree to which the person sees the outcome in question as leading to the attainment of other outcomes. Instrumentality varies from -1 (meaning that the outcome in question is perceived as always leading to not attaining the second outcome) to $+1$ (meaning that the outcome is perceived as always leading to the attainment of the second outcome).

Although this model can be used to predict the valence of any outcome, it has been applied most frequently to the prediction of job satisfaction. In essence, the model says that the worker's satisfaction with his job results from the instrumentality of the job for attaining other outcomes and the valence of those outcomes. In the remainder of this section we refer to this model as the *job satisfaction model.*

Vroom's second model predicts the force toward behavior. The force on a person to perform an act is conceptualized by Vroom as a monotonically increasing function of the algebraic sum of the products of the valences of all outcomes, and the strength of his expectancies that the act will be followed by the attainment of these outcomes (Vroom, 1964). Symbolically,

$$F_i = \sum_{j=1}^{n} (E_{ij} V_j)$$

where

F_i = the force on the individual to perform act i;

E_{ij} = the strength of the expectancy that act i will be followed by outcome j;
V_j = the valence of outcome j;
n = the number of outcomes.

The individual's expectancy is defined by Vroom as his belief concerning the probability that the behavior in question will be followed by the outcome of interest. An expectancy is a perceived probability, and therefore ranges from zero to plus one. It is distinguished from instrumentality in that it is an action-outcome association, while instrumentality is an outcome-outcome association. While expectancies are perceived probabilities, instrumentalities are perceived correlations.

Vroom suggested that this second model can be used to predict choice of occupation, remaining on the job, and effort. In practice, it has been tested with respect to job performance, and we, therefore, refer to it as the *job performance model*. Specifically, Vroom stated that the force on the individual to exert a given amount of effort is a function of the algebraic sum of the products of the valence of each level of performance and the person's expectation that each level of performance will be attained by that amount of effort. Note that the amount of effort, not performance, is predicted by Vroom; effort is considered to be a behavior, while performance is an outcome.

Galbraith and Cummings (1967) elaborated and tested Vroom's job performance model. They began by distinguishing between first- and second-level outcomes. A first-level outcome is one which has a valence that the investigator is interested in predicting—in their case, performance on the job. They defined second-level outcomes as events to which the first-level outcomes are expected to lead. It should be noted that Galbraith and Cummings' definition of second-level outcomes suggests that only those outcomes actually expected by the individual to result from the first-level outcome may be considered. In fact, outcomes not perceived by the individual as resulting from the first-level outcome may be included in the equation, since their instrumental value will be zero; and they will, therefore, not degrade the prediction of the valence of the first-level outcome.

Galbraith and Cummings described three sources of second-level outcomes that might be expected to result from performance. First, the organization as a whole may provide money, fringe benefits, and promotions for performance. Second, the supervisor may be supportive. Third, the work group may be accepting. Essentially, Galbraith and Cummings are providing an a priori list of the outcomes they feel are important to the prediction of the valence of high performance. The problem with this approach is that the investigators may fail to support the instrumentality hypothesis because they choose the second-level outcomes that are not important to the subjects, rather than because of any inadequacy of the hypothesis.

Galbraith and Cummings did not study the effects of expectancy in this investigation. Rather they attempted to predict performance for a sample of subjects considered likely to have expectancies of + 1 about the relationship between their behavior and high performance. Workers were chosen whose attainment of performance did not depend on other workers and whose work pace was not dependent on a machine. Their sample consisted of 32 workers in an Indiana heavy equipment plant. Measures of instrumentality of performance for each of the second-level outcomes and the valence of each second-level outcome were obtained from subjects. The measure of performance was obtained from company records. This measure took the form of a percentage of the company standard for a 1-month period.

The instrumentality hypothesis was tested separately for each of the second-level outcomes. Significant prediction of performance was obtained only for the outcome, supportiveness of the supervisor. Neither the valence of this outcome nor the instrumentality of performance for the outcome was related to performance by itself, but the interaction of the two was related to performance. This result is as Vroom's job performance model predicts.

The instrumentality hypothesis was also tested for all outcomes taken together. Stepwise multiple regression was used. In this analysis, the interaction of valence and instrumentality was significantly related to performance for both the supportiveness and money outcomes. It should be noted that the analysis of these data did not include a test of the association between the *sum* of the products of instrumentality and valence ratings. This sum was central to Vroom's statement of the instrumentality hypothesis, but Galbraith and Cummings (1967) gave no explanation for the absence of this test.

Despite the author's conclusions, we do not feel these results provide impressive support for instrumentality theory. The failure to obtain strong support of the instrumentality hypothesis may be due to a number of factors besides the inadequacy of the hypothesis. First, subjects' expectancies that their behavior leads to high performance may not have been controlled at + 1 as the authors hoped. According to Vroom's hypothesis, even when high performance is positively valent, the individual will not exert a given amount of effort if he does not *expect* his effort to lead to high performance. Thus, if subjects' expectancies were not at + 1, the theory would not predict that the sum of the products of the valence of second-level outcomes multipled by the instrumentality of performance for those second-level outcomes would be related to performance.

Second, the a priori set of outcomes used in this study may have excluded outcomes related to performance. For example, a feeling of accomplishment may be an important outcome of achieving high performance for some individuals. Negative outcomes, such as getting fired and being bored, may also affect performance. For example, some individuals may perceive many positively valued outcomes as resulting from performance, but also perceive boredom as a result. If boredom is strongly negatively valenced, it may lead the individual to forego the possibility of attaining positive outcomes.

A third reason for the failure of Galbraith and Cummings' (1967) study to support the theory is that amount of effort and performance may not have been related in this sample. Recall that Vroom's (1964) model predicts effort, not performance. Thus, if Galbraith and Cummings had measured workers' intentions to exert effort or actual effort exerted, they might have found that the model adequately predicted these variables.

Lawler and Porter (1967) did attempt to predict *effort* from subjects' perceived instrumentalities. They studied 154 managers in five different organizations, and ratings of the importance of seven outcomes were collected: pay, promotion, prestige, security, autonomy, friendship, and opportunity to use skills and abilities. A composite measure of the subject's expectancies was based on the sum of his ratings of the degree to which (*a*) effort, (*b*) high productivity, and (*c*) good job performance led to the seven outcomes. Ratings of effort and performance were obtained from superiors, peers, and the subject. Multiple correlations between the Expectancy × Importance ratings for each of the seven outcomes, and the supervisor, peer, and self-ratings of effort, were .27 ($p < .10$), .30 ($p < .05$), and .44 ($p < .01$), respectively. Prediction of performance from the Expectancy × Outcome scores was less successful; the multiple correlations were .18 (*ns*), .21 (*ns*), and .38 ($p < .01$) for the supervisor, peer, and self-performance ratings. With seven predictors, these multiple correlations are hardly strong support for the instrumentality hypothesis. As in the Galbraith and Cummings (1967) study, Lawler and Porter's use of multiple correlations is not consistent with Vroom's (1964) model (nor with their own statement of the model). A simple sum of the Importance × Expectancy scores would have been the appropriate predictor. Moreover, the authors used an a priori set of outcomes. The problem with this approach was discussed above.

A study by Lawler (1968) provides support for the causal status of subjects' instrumentalities. Six rewards were rated by 55 managers as to their importance. Subjects' expectations of the degree to which effort and job performance led to each of the six rewards was also rated. Peer, supervisor, and self-ratings of performance were obtained at the same time. All of these measures were obtained at the beginning and end of a 1-year interval. Multiple correlations were generated to assess the ability of the six Expectancy × Importance scores to predict performance. For the supervisors' ratings, these correlations were .44 (Time 1), .52 (Time 2), .55 (Time 1 Expectancy × Importance with the Time 2 Performance), and .39 (Time 2 Expectancy × Importance with the Time 1 Performance). The first three coefficients were significant ($p < .05$). The corresponding correlations for peer ratings were .52, .30, .45, and .30 ($p < .05$ for the first and third coefficients), and for the self-performance ratings they were .43, .48, .65, and .52 ($p < .05$ for the last three). The causal status of the expectancy attitudes is supported by the fact that Time 1 Expectancy × Importance measures are more highly related to Time 2 performance measures than Time 2 Expectancy × Importance measures are related to Time 1 performance. Again, however, multiple correlation with six Expectancy × Importance predictors was used, and the rewards were not generated by the subjects themselves.

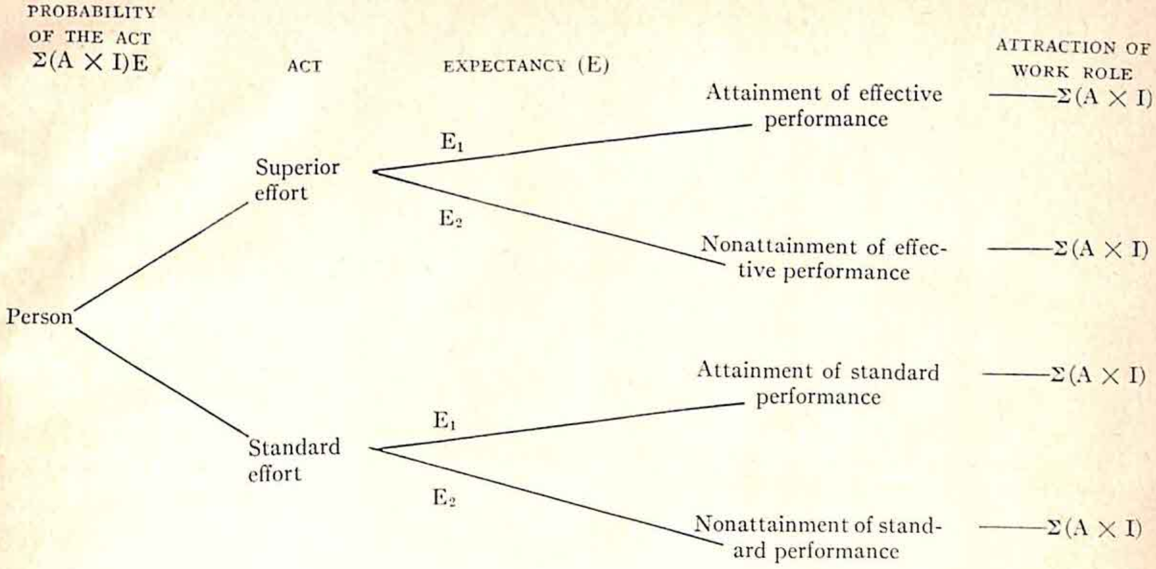
Another experiment by Hackman and Porter (1968) avoided some of the problems with the above studies. These authors used instrumentality theory to predict the effectiveness of 82 telephone company service representatives. Instead of developing their own outcomes, they had 24 service representatives generate relevant outcomes. The 14 most frequently mentioned outcomes were included, as well as 4 negative outcomes which they added. The 82 subjects were then asked to evaluate each outcome and the probability that working hard on the job would lead to each outcome. The sum of the products of expectations and valence was then correlated with 11 effectiveness measures. The correlations ranged from .25 to .40, and 9 of the 11 were significant ($p < .05$). These results suggest that instrumentalities are important, but not the only factor related to employee performance.

Graen (1969) tested Vroom's job performance and job satisfaction models and examined some factors affecting perceived instrumentality. His results define important boundary conditions for instrumentality theory. In addition, he proposed some useful modifications and extensions of the theory.

Graen began by clarifying the distinction between first- and second-level outcomes. A first-level outcome is called a work role. A work role is defined as a set of behaviors expected by the organization and considered appropriate for the occupant of a position in the organization. For example, "effective job performer," "team member," and "group leader" are work roles. Work roles are attained and maintained through the performance of the expected behaviors in such a way that minimal organizational expectations are met. Second-level outcomes are called role outcomes. They are the outcomes that accrue to the individual as a result of his attainment or maintenance of the work role. Graen focused on two work roles: *effective performer* and *job incumbent*. Eight role outcomes (second level) are included in his study: accomplishment, achievement feedback, salary, human relations, recognition, policies and practices, responsibility, and working conditions. As in the Galbraith and Cummings (1967) study, the role outcomes were derived on an a priori basis.

Figure 1 presents the job satisfaction and job performance models of Graen (1969). According to him, job satisfaction is a matter of satisfaction with the job-incumbent work role. It is predicted to be a monotonically increasing function of the sum of the products of the instrumentality of the job-incumbent work role for attaining role outcomes and the valence of those role outcomes. Job performance, on the other hand, is predicted to be a positive function of the valence of the effective performer work role and the perceived expectancy that increased effort will lead to attainment of this work role.

The study was an experimental simulation in which women were hired for part-time, temporary employment by a fictitious company. In the first task, subjects had to find certain correlations from a 155 × 155 matrix and transcribe them. The second task was the same as the first, except that the correlations were to be rounded to two places before transcription. There were three treatment groups. The first, the achievement group, was rewarded for prior effective performance. The second group, the incentive group, was offered money as an inducement for subsequent performance. The third group was a control group: the money they obtained was not contingent on performance. Subjects were run during two sessions, and the treatments were introduced in the second session. These treatments were intended to permit a test of the boundary conditions of the instrumentality theory and to examine the factors affecting perceived instrumentality. Upon completion of the task


PROBABILITY OF THE ACT $\Sigma(A \times I)E$
ACT
EXPECTANCY (E)
ATTRACTION OF WORK ROLE
Person
Superior effort
Standard effort
E_1
E_2
E_1
E_2
Attainment of effective performance
Nonattainment of effective performance
Attainment of standard performance
Nonattainment of standard performance
$\Sigma(A \times I)$
$\Sigma(A \times I)$
$\Sigma(A \times I)$
$\Sigma(A \times I)$

where attraction of work role equals Graen's extended model of work role attraction (job satisfaction):

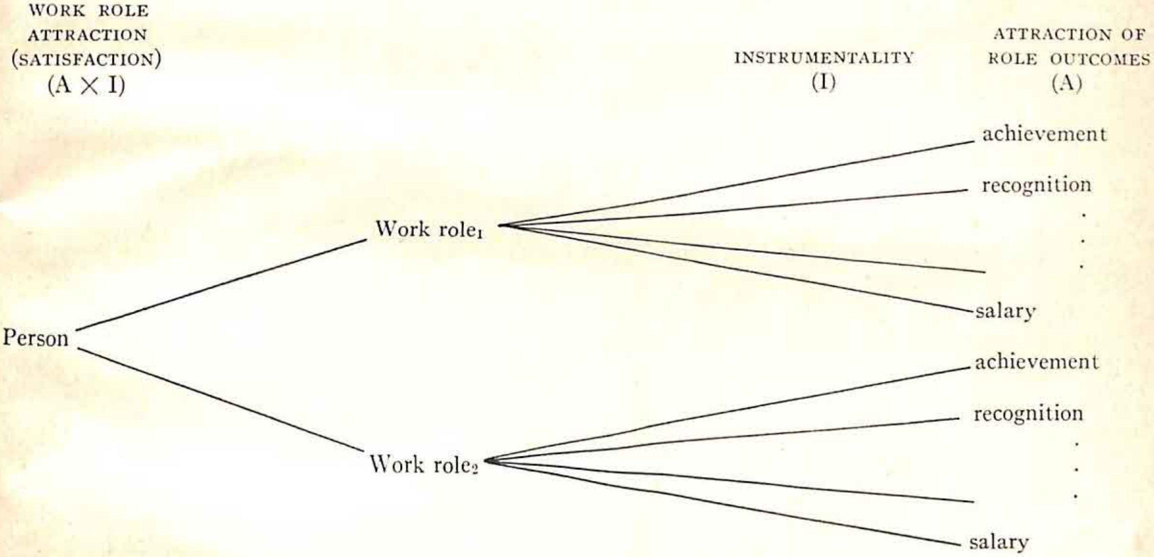


FIG. 1. Graen's extended model of the probability of an act (job performance).

activity, measures of the following were obtained: the attraction of work-role outcomes, the perceived instrumentality of the two work roles (the particular job and "effective performer") for the attainment of each of the *eight* role outcomes, and the perceived expectancy that increased effort would lead to more effective performance.

A separate job satisfaction measure based on the Hoppock Job Satisfaction Blank (Hoppock, 1935) was used to measure job satisfaction. Performance measures used were the number of items completed (quantity) and proportion correct (quality). Both a raw score and a gain score (pre- to posttreatment) for both quantity and quality measures were used in the final analysis.

Perceived instrumentality varied according to the treatment condition. It will be recalled that the achievement treatment group was rewarded for prior effective performance and that the incentive treatment group received money as an inducement for subsequent good performance. For the achievement treatment group, the perceived instrumentality of the job-incumbent work role for achievement and recognition was significantly higher than it was for the incentive and control groups. For the incentive treatment group, the perceived instrumentality of the job-incumbent work role for money was significantly higher than it was for the other two groups. Graen (1969) also looked at the perceived instrumentality of the effective performer work role. The achievement treatment group rated the instrumentality of this work role for achievement and recognition significantly higher than did the groups in the other two treatment conditions. However, the perceived instrumentality of the effective performer work role for money was *not* greater for the incentive treatment group.

The results of this study provide partial support for the job satisfaction model and suggest some limitations for it. The specific model tested in this study states that satisfaction with the work role of job incumbent is a monotonically increasing function of the sum of the products of the valences of role outcomes and the perceived instrumentality of the work role for these outcomes. The model was tested separately for intrinsic outcomes (accomplishment, achievement feedback, recognition, and responsibility), and extrinsic outcomes (human relations, policies and practices, salary, and working conditions), and for all outcomes taken together. It was found that the model predicted both job satisfaction (i.e., the Hoppock scale scores) and gains in satisfaction for the achievement-feedback treatment group when intrinsic outcomes and all outcomes were considered. These correlations were .36 (intrinsic outcomes) and .37 (all outcomes) for the raw scores, and .32 and .29 for the gain scores ($p < .05$ for all coefficients). Since the achievement-feedback group was the only treatment group for which a contingency between the work role and certain role outcomes was clearly stated, Graen concluded that the job satisfaction model is limited to situations where the contingencies between work roles and role outcomes are stated in a concrete manner.

The job performance model (effective performer) failed to predict either the quality or quantity raw performance scores. In other words, there was no correlation between the raw scores and the product of the expectation that effort would lead to the role of effective performer and the evaluation of that role. However, the model was partially successful in predicting residual gain scores. These scores were obtained by partialing out pretreatment performance scores so that the variance in posttreatment performance due to ability, and pretreatment motivation is minimized. The effective performer model significantly predicted these scores for the achievement-treatment group when intrinsic and extrinsic role outcomes were considered separately and when all outcomes were considered at the same time. These correlations were .29, .22, and .28 for one task, and .41, .31, and .39 for another ($p < .05$ for all six coefficients). The model did not predict residual gain scores for the other treatment groups.

Together with the job satisfaction results, the just-mentioned results define boundary conditions for the instrumentality theory models. Graen suggested that the theory will be successful only when the contingencies between work roles and role outcomes are clearly stated.

These data also suggest, though less clearly, that the models will only be successful for the specific role outcomes that are linked to the work role. This conclusion is based on the fact that the job satisfaction model successfully predicted job satisfaction only for the intrinsic outcomes, and these outcomes were the only outcomes specifically linked to the job. Moreover, in the case of the relationship between the job performance model and performance for the achievement group, the correlations were higher for the intrinsic outcomes than they were for the extrinsic outcomes.

Extended Models

Both Graen (1969) and Porter and Lawler (1968) suggested extensions of the basic instrumentality model to predict performance and job satisfaction. The latter authors suggested that the effort-performance relationship is moderated by ability and role perception variables. The relevant role perceptions are concerned with the types of effort an individual believes are essential to effective job performance. They argued that an accuracy of role perceptions score should be multiplied by an Effort × Ability score to predict performance. They presented some evidence that role perceptions do indeed moderate the effort-performance relationship (Lawler & Porter, 1967; Porter & Lawler, 1968), but no data have been presented that measure all three variables and combine them in the suggested manner.

Graen proposed a number of modifications of the models in addition to the distinction between work roles (first-level outcomes) and role outcomes (second-level outcomes). First, Graen suggested that the focus of the model be on gains in performance or satisfaction rather than raw scores. Besides being consistent with his results, he argued, this focus makes the model a dynamic one and provides the model with developmental implications. It is not clear why the prediction of changes in performance or satisfaction is more important than the prediction of raw scores, but Graen's modified model, if successful, would make explicit the factors that are important for increasing production or satisfaction.

Graen (1969) suggested additional predictors of performance and satisfaction. In addition to the Expectancy × Valence term, gains in performance are predicted from (*a*) the perceived expectations of others and pressure to comply with the others and (*b*) the sum of the valence of intrinsic consequences of the performance gain multiplied by the expectancy that the gain will lead to these consequences. The latter term reflects the subject's internally felt pressure toward increased performance.

Gains in satisfaction that are associated with gains in performance are predicted from the same model, except that the term referring to others' expectations is the amount of pressure consonant with increased performance *relative* to all pressure and the intrinsic instrumentality term refers to the amount of internally felt pressure that is consonant with increased performance relative to all internal pressure.

It can be seen that both Porter and Lawler's and Graen's work are conceptually similar to that of Fishbein and Dulany in that they include other sources of influence on behavior besides instrumentalities. Empirical support for these extensions, however, still must be presented.

In summary, instrumentalities appear to be an important factor in predicting the satisfaction and behavior of personnel in organizational settings. However, if we compare the amount of variance accounted for in the organization studies with the amount accounted for in the attitude and verbal conditioning areas, it appears that the use of instrumentalities has been less successful in the former area. There are at least two reasons why this might be the case. First, the phenomena that the industrial psychology studies are attempting to predict are more complex than those predicted in the various game and verbal conditioning settings. Second, the situations in the industrial settings are less controllable than those in the attitude and verbal conditioning areas.

Discussion and Evaluation

The evidence reviewed in this study provides fairly good support for the basic instrumentality hypothesis. The hypothesis that behavior is related to the perception that it will lead to certain outcomes and the valence of those outcomes received some support in three areas of psychology: verbal conditioning, attitudes, and industrial psychology. This section discusses the strengths and weaknesses of the theories that have been reviewed in this study. For the most part, these theories contain the instrumentality hypothesis and have additional components referring to the normative influences on behavior. We begin by pointing out some of the favorable aspects of these theoretical approaches.

1. Both the instrumentality hypothesis and the more extended theories that contain it provide a systematic framework for the addition of new variables. One might isolate additional outcomes that are important for behavior in particular situations. Similarly, one might have different normative components corresponding to the perceived expectations of different groups in the subject's role set, for example, the peer and management groups in the worker's role set. Moreover, in the formulations of Dulany, Fishbein, and Graen entirely new components can be added if there

is reason to believe that they affect behavior and behavioral intention.

2. The theories reviewed here represent a major step toward the treatment of the complexity of social behavior. For many years, psychologists have related variables in a pairwise fashion (e.g., need achievement and production), with the net result that we have many specific relations and very little integration. The theories reviewed above provide a structure allowing the psychologist to examine the effects of a number of variables that may be simultaneously affecting behavior. The instrumentality component incorporates a whole set of outcomes and probabilities that may influence a given behavior. Dulany's, Fishbein's, and Graen's theories include the expectations of a set of other people and one's willingness to comply with these expectations along with instrumentalities.

3. The theories described in this study suggest important sources of behavioral change. The instrumentality component suggests two ways in which behavior may be changed. First, the subject's perception of the instrumental relations between behavior and its outcomes may be changed. For example, employees may value the rewards provided by the organization but not know which behaviors lead to them. Second, behavior may be influenced by changing the value of outcomes or organizational rewards. These theories also suggest that changes in the behavior of the subject result from the expectations of others. Thus, one might also bring about changes in the target subject by changing the expectations of those around him.

4. Finally, in the first section we discussed the necessity of using a logic of theoretical networks in order to support the contention that unobservable constructs are causes of behavior. This logic requires richly articulated theories including a network of constructs and operations. The instrumentality theories that include a normative component and, to a lesser extent, the instrumentality idea by itself are relatively good examples of how this approach to theory building could be used. These theories, as pointed out above, (*a*) can use a large number of variables, (*b*) are open to new additions, and (*c*) have received relatively good support when compared to other interpretations (e.g., the comparison with experimental manipulations by Dulany).

There are also a number of problems associated with this theoretical approach and with the statistical techniques typically employed to test it.

1. The theories demand specific rules of combination that exclude the possibility of examining certain interaction effects. Two outcomes, when they are both perceived as possible consequences of an act, may contribute more to the intentions of an individual than when they appear separately with respect to other behaviors. That is, outcome combinations may be important. To date, this possibility does not appear to have been tested.

2. Many of the theories that we reviewed above are tested by multiple regression. In their most common use, multiple regression equations limit the investigator to the same weighted linear function as the rule of combination of components for all subjects. This constraint is not required by the theory, and, in fact, it was suggested that the relative weight of each factor may differ over individuals and situations. A weighting system that is psychologically meaningful in the sense that it reflects the way in which individuals actually combine these factors is needed.

3. Even when the theories are able to predict behavioral intentions or behavior with excellent accuracy, there appear to be situations where behavioral intentions or instrumentalities are not highly related to actual behavior. Fishbein (1967b) has pointed out that the degree of control that the individual has over his behavior is important. In some cases, one's behavior may be controlled by another person as, for example, when an employee intends to see his boss. These theories must either include variables that mediate the behavioral intention–behavior relationships or limit themselves to behaviors the individual does control.

4. It was mentioned earlier that the constructs in the instrumentality component change over time. That is, perceived instrumentalities and the value of outcomes often change; so do the perceived expectations of others. When this change cannot be predicted or controlled, behavioral intentions and the

components of the theory may not be related to actual behavior.

5. The preceding three criticisms point to the need for a theory of the situation. Such a theory should describe those aspects of the situation that affect the relative weights of the factors in the instrumentality equation and that mediate the relationship between behavioral intention and behavior. The authors and their colleagues are presently attempting to develop a structural role theory (Oeser & O'Brien, 1967; Oeser & Harary, 1962, 1964) approach to the factors affecting the weights of the components in the instrumentality equation.

6. Many of the theories discussed above use a standard set of outcomes for all subjects. It is likely that different outcomes are relevant for different subjects. Moreover, Rosenberg's (1956) data suggest that attitude is linked to only a small number of outcomes and that additional outcomes degrade prediction. This suggests that subjects should be asked to generate their own outcomes.

7. Finally, we are struck by the possibility that the instrumentality component of these theories may be culturally specific. Thus, individuals who think solely in terms of the outcomes accruing as the result of their behavior may not be found in cultures other than our own. Similarly, there may be sex differences in the applicability of the instrumentality hypothesis.

In summary, instrumentality theories are being used in a wide range of areas in psychology. Theories using these concepts have generally been supported. Additional work with these theories should integrate them with a theory of the situation, develop psychologically meaningful weighting systems for the factors included, and examine the interaction of factors as they affect behavioral intention and behavior.

REFERENCES

AJZEN, I., & FISHBEIN, M. The prediction of behavioral intentions in a choice situation. *Journal of Experimental Social Psychology*, 1969, **5**, 400–416.

BEACH, L. R., & WISE, J. A. Subjective probability and decision strategy. *Journal of Experimental Psychology*, 1969, **79**, 133–138. (a)

BEACH, L. R., & WISE, J. A. Subjective probability estimates and confidence ratings. *Journal of Experimental Psychology*, 1969, **79**, 438–444. (b)

BECKER, G. M., & MCCLINTOCK, C. G. Value: Behavioral decision theory. *Annual Review of Psychology*, 1967, **18**, 239–287.

CARNAP, R. The methodological character of theoretical concepts. In H. Feigl & M. Scriven (Eds.), *Minnesota studies in the philosophy of science*. Vol. 1. Minneapolis: University of Minnesota Press, 1956.

CRONBACH, L., & MEEHL, P. Construct validity in psychological tests. *Psychological Bulletin*, 1955, **52**, 281–302.

DOOB, L. W. The behavior of attitudes. *Psychological Review*, 1947, **54**, 135–156.

DULANY, D. E. Hypotheses and habits in verbal "operant conditioning." *Journal of Abnormal and Social Psychology*, 1961, **63**, 251–263.

DULANY, D. E. The place of hypotheses and intentions: An analysis of verbal control in verbal conditioning. In C. Eriksen (Ed.), *Behavior and awareness*. Durham: Duke University Press, 1962.

DULANY, D. E. Awareness, rules, and propositional control: A confrontation with S–R behavior theory. In D. Horton & T. Dixon (Eds.), *Verval behavior and general behavior theory*. Englewood Cliffs, N. J.: Prentice Hall, 1968.

EDWARDS, W. Behavioral decision theory. *Annual Review of Psychology*, 1961, **12**, 473–498.

EDWARDS, W. Subjective probabilities inferred from decisions. *Psychological Review*, 1962, **69**, 109–135.

FEIGL, H. Some major issues and developments in the philosophy of science of logical empiricism. In H. Feigl & M. Scriven (Eds.), *Minnesota studies in the philosophy of science*. Vol. 1. Minneapolis: University of Minnesota Press, 1956.

FISHBEIN, M. An investigation of the relationship between beliefs about an object and the attitude toward that object. *Human Relations*, 1963, **16**, 233–240.

FISHBEIN, M. Prediction of interpersonal preferences and group member satisfaction from estimated attitudes. *Journal of Personality and Social Psychology*, 1965, **1**, 663–667. (a)

FISHBEIN, M. A consideration of beliefs, attitudes, and their relationships. In I. D. Steiner & M. A. Fishbein (Eds.), *Current studies in social psychology*. New York: Holt, Rinehart & Winston, 1965. (b)

FISHBEIN, M. A consideration of beliefs and their role in attitude measurement. In M. Fishbein (Ed.), *Readings in attitude theory and measurement*. New York: Wiley, 1967. (a)

FISHBEIN, M. Attitude and the prediction of behavior. In M. Fishbein (Ed.), *Readings in attitude theory and measurement*. New York: Wiley, 1967. (b)

GALBRAITH, J., & CUMMINGS, L. L. An empirical investigation of the motivational determinants of task performance. Interactive effects between instrumentality-valence and motivation-ability. *Organizational Behavior and Human Performance*, 1967, **2**, 237–257.

GEORGOPOULOS, B. S., MAHONEY, G. M., & JONES, N. W., JR. A path-goal approach to productivity. *Journal of Applied Psychology*, 1957, **41**, 345–353.

GRAEN, G. Instrumentality theory of work motivation:

Some experimental results and suggested modifications. *Journal of Applied Psychology Monograph*, 1969, **53**, (2, Pt. 2).

HACKMAN, J. R., & PORTER, L. W. Expectancy theory predictions of work effectiveness. *Organizational Behavior and Human Performance*, 1968, **3**, 417–426.

HEMPEL, C. G. *Fundamentals of concept formation in empirical science*. Chicago: University of Chicago Press, 1952.

HOPPOCK, R. *Job satisfaction*. New York: Harper and Brothers, 1935.

LAPIERE, R. T. Attitudes versus actions. *Social Forces*, 1934, **13**, 230–237.

LAWLER, E. E. A correlational-causal analysis of the relationship between expectancy attitudes and job performance. *Journal of Applied Psychology*, 1968, **52**, 462–468.

LAWLER, E. E., & PORTER, L. W. Antecedent attitudes of effective managerial performance. *Organizational Behavior and Human Performance*, 1967, **2**, 122–142.

LEWIN. K. *A dynamic theory of personality*. New York: McGraw-Hill, 1935.

LIKERT, R. A technique for the measurement of attitudes. *Archives of Psychology*, 1932, No. 140, 44–53.

LOCKE, E. A., BRYAN, J. F., & KENDALL, L. M. Goals and intentions as mediators of the effects of monetary incentives on behavior. *Journal of Applied Psychology*, 1968, **52**, 104–121.

MALTZMAN, I. Awareness: Cognitive psychology versus behaviorism. *Journal of Experimental Research in Personality*, 1966, **1**, 161–165.

OESER, O. A., & HARARY, F. A mathematical model for structural role theory. I. *Human Relations*, 1962, **15**, 89–109.

OESER, O. A., & HARARY, F. A mathematical model for structural role theory. II. *Human Relations*, 1964, **17**, 3–17.

OESER, O. A., & O'BRIEN, G. E. A mathematical model for structural role theory. III. The analysis of group tasks. *Human Relations*, 1967, **20**, 83–97.

OSGOOD, C. E., SUCI, G. J., & TANNENBAUM, P. H. *The measurement of meaning*. Urbana, Ill.: University of Illinois, 1957.

PEAK, H. Attitude and motivation. *Nebraska Symposium on Motivation*, 1955, **3**, 149–188.

PETERSON, C. R., & BEACH, L. R. Man as an intuitive statistician. *Psychological Bulletin*, 1967, **68**, 29–46.

PETERSON, C. R., & MILLER, A. J. Sensitivity of subjective probability revision. *Journal of Experimental Social Psychology*, 1965, **70**, 117–121.

PORTER, L. W., & LAWLER, E. E. *Managerial attitudes and performance*. Homewood, Ill.: Irwin-Dorsey, 1968.

RAPOPORT, H. & CHAMMAH, A. M. *Prisoner's Dilemma: A study on conflict and cooperation*. Ann Arbor: University of Michigan Press, 1965.

ROSENBERG, M. J. Cognitive structure and attitudinal affect. *Journal of Abnormal and Social Psychology*, 1956, **53**, 367–372.

SPIELBERGER, C. D. Theoretical and epistemological issues in verbal conditioning. In S. Rosenberg (Ed.), *Directions in psycholinguistics*. New York: Macmillan, 1965.

TOLMAN, E. C. *Purposive behavior in animals and men*. New York: Appleton-Century, 1932.

TRIANDIS, H. C. Exploratory factor analyses of the behavioral component of social attitudes. *Journal of Abnormal and Social Psychology*, 1964, **68**, 420–430.

TRIANDIS, H. C. *Attitudes and attitude change*. New York: Wiley, 1970.

VROOM, V. H. *Work and motivation*. New York: Wiley, 1964.

(Received May 19, 1970)

ATIONAL AND PSYCHO
LIBRARY

Manuscripts Accepted for Publication in the

Psychological Bulletin

The Septum and Behavior: A Review: P. A. Fried*: Department of Psychology, Carleton University, Ottawa 1, Canada.

Retrieval Rules and Semantic Coding in Short-Term Memory: A. D. Baddeley*: Laboratory of Experimental Psychology, University of Sussex, Brighton, England BN1 9QY.

Note on "Factors Influencing the Outcome of Psychotherapy": H. J. Eysenck*: Institute of Psychiatry, University of London, Bethlehem Royal Hospital, London S. E. 5.

Another Reply to Eysenck: Lester Luborsky*: University of Pennsylvania, School of Medicine, Universit Hospital, 207 Piersol Building, 3400 Spruce Street, Philadelphia, Pennsylvania 19104.

Molecular Changes during Learning—Behavioral Strategy: A Comment on Gaito and Bonnet: William T. Gree ough* and Steven F. Maier: Department of Psychology, University of Illinois at Urbana-Champaign, Cha paign, Illinois 61820.

Controls in Molecular Psychobiological Research: John Gaito* and Kenneth Bonnet: Department of Psycholog York University, 4700 Keele Street, Downsview 463, Ontario.

The Hippocampus and Motivation: Leonard E. Jarrard*: Department of Psychology, Washington and Le University, Lexington, Virginia 24450.

The Efficacy of Psychiatric Rehabilitation: William A. Anthony,* Sara Sharratt, Gregory J. Buell, and Michae Althoff: Department of Psychology, Southern Illinois University, Carbondale, Illinois 62901.

The Hudson-Dunn Clustering Index Revisited: Robert L. Hudson*: Department of Psychology, Mississipp State University, P. O. Drawer PF, State College, Mississippi 39762.

---

* Asterisk indicates author for whom the address is supplied.

REHABILITATION INDIA
CALCUTTA
700068